ATTENDING TO THE POSITIVE: RETROSPECTIVE VALIDATION OF THE SAPROF-SO

By Thomas Nolan

A thesis submitted in partial fulfillment of the requirements for the Degree of Master of Science in Psychology

Department of Psychology

University of Canterbury

2021

Acknowledgements

Thank you to my co-supervisors Gwenda Willis and Sarah Beggs Christofferson for your seemingly endless patience, guidance, and support during the production of this thesis. Without your vast knowledge and unwavering willingness to decipher poorly written drafts, this thesis would never have come to fruition. Further, thank you for granting me the opportunity to conduct this research and trusting me to complete it. In addition, special thanks to Gwen for taking the time to fly down to Christchurch to code the required cases for the interrater reliability analyses.

Thank you to Fay McIntyre for swiftly and expertly coordinating the retrieval of what felt like millions of case files. The mountain of administrative work that you completed will never be forgotten. Thanks to the staff at Kia Marama STU for kindly accommodating me in your workspace, sporadically and without warning, for two and a half years. The Department of Corrections must also be acknowledged for approving this study and maintaining the database that makes this type of research possible.

Thanks to my parents, Jim and Janeen, for your moral support, and more financial support than I am comfortable admitting to. Further, thanks to Harriet, Ben, Sam, Matt, and my fellow psychology students for tolerating the ramblings of a tired and increasingly cynical postgraduate student. Finally, thanks to Samyang and Nongshim for manufacturing the delicious yet reasonably priced instant noodles that provided me with the sustenance to complete this thesis.

Table of Contents

List of Tables

Abstract

Forensic risk assessment tools almost exclusively measure factors that are associated with increased rates of recidivism, meaning factors that are associated with decreased rates of recidivism have historically been ignored, leaving risk assessment unbalanced. The present study is the first to investigate the predictive validity of the Structured Assessment of Protective Factors for violence risk – Sexual Offence version (SAPROF-SO), a new and unique tool designed to assess hypothesised protective factors against sexual recidivism in adult males. The study had a further aim of assessing the incremental validity of the SAPROF-SO over and above a well-validated measure of static risk (Static-99R). In this retrospective study, SAPROF-SO scores were informed by analysing archived case files of 210 men with convictions for child sexual offences. To minimise item omissions and maximise interrater reliability, a supplementary scoring guide theoretically aligned with the official manual was created and utilised. SAPROF-SO Total and Personal domain scores were significantly predictive of sexual recidivism after a long follow-up period (M = 12.24 years) with AUC values between .78 and .81. Total and Personal scores were also significantly predictive of violent and general recidivism, but to a lesser extent. Professionally Provided Support domain scores were not significantly predictive of any type of offending. Total and Personal scores additionally provided significant incremental validity over and above Static-99R scores in the prediction of sexual recidivism. Results of the current study provide rationale for the inclusion of protective factors in risk assessment and invite further research validating this tool.

Attending to the Positive: Retrospective Validation of the SAPROF-SO

Based on results from a survey of almost seven thousand people, The Ministry of Justice (2016a) estimated that in 2013, 186,000 sexual offences occurred in New Zealand. Further, they conceded this statistic is likely an underestimation. A small yet noteworthy proportion of these offences were likely committed by people who had been released following imprisonment for sexual crime. Nadesu (2011) analysed New Zealand data from a group of 1100 men imprisoned for sexual offences and found that 10% of those convicted for sexual offences against adults were reimprisoned for another sexual offence within 5 years of release, and this statistic was 5% for those originally imprisoned for sexual offences against children. Because it is estimated that only 9% of sexual offences are reported to police (Ministry of Justice, 2016b), and some of the sample may have accrued further sexual charges and convictions that did not result in reimprisonment, these statistics likely only represent a fraction of the total recidivism within this group.

The solution to sexual recidivism may partially reside in improving risk assessment for people convicted of sexual crime. Risk assessment is of crucial importance to the criminal justice system. Information from these evaluations influence important decisions that can have significant implications for those imprisoned, their families, their victims, and wider society (Beggs & Grace, 2010; Doren, 2006). In forensic psychology literature, risk assessment is defined as the process of estimating the likelihood that an individual will reoffend (Beggs, 2008; Cording & Beggs Christofferson, 2017). However, modern assessments yield more than a mere probability of reoffending, they also provide a wealth of relevant and important information. Information from these evaluations affect sentencing, security ratings, who is assigned to limited treatment programmes, what is targeted in treatment, whether or when to release, and probationary conditions (Beggs, 2008). By improving risk assessment, institutions will be able to better allocate individuals to suitably

intense and individualised (and therefore more efficacious; Andrews, 1995) treatment and decisions could be made more appropriately regarding incarceration and reintegration into society. This could reduce recidivism, prevent individuals from facing unnecessarily severe sanctions, and lead to more efficient spending of taxpayer money (Beggs, 2008; Cording & Beggs Christofferson, 2017).

The introduction begins with a brief history of risk assessment methodology and explanation of key terminology. This leads into a discussion of protective factors and the proposed benefits of incorporating them into risk assessment. Next, the tool at the focus of the current study - the Structured Assessment of Protective Factors for violence risk – Sexual Offence version (SAPROF-SO) is introduced, followed by a discussion of the reliability and validity of this and related measures, and an overview of the current study.

Risk Assessment Methodology and Terminology

To best understand risk assessment, it is useful to grasp how it has evolved into its current form. Bonta (1996) divides the history of the prediction of criminal behaviour into three generations. The first generation was Professional Judgement or Unstructured Professional Judgement (UPJ). Assessment was originally undertaken by having experts in the social sciences interview individuals before reaching a conclusion on an individual's likelihood of reoffending in an unstructured manner (Bonta & Andrews, 2017). Methods between different experts could vary greatly and therefore the predictive accuracy of these judgements would follow suit. Assessors varied in the psychological, environmental and historical factors they deemed important, and the psychometric tests they administered. The opinions of these clinicians were reportedly often not rooted in empirical science and were guided by "gut feelings", intuition and other unobservable criteria (Bonta & Andrews, 2017, p. 193). Andrews et al. (2006) combined data from three meta-analyses that investigated the accuracy of UPJ and found that this methodology had only weak predictive validity.

The second generation of risk assessment was defined by the introduction of actuarial static risk scales (Bonta, 1996; Bonta & Andrews, 2017). These scales (or tools as they are otherwise known) represent a more structured, mathematical, and empirical approach to assessing risk. Actuarial risk assessment tools include multiple "items" which are informed by empirical research identifying variables that have been shown to be associated with an increase in the likelihood that a person will reoffend upon release, such as young age or history of abuse. These variables are known as "risk factors". These tools have scoring criteria for each item and a scoring form on which the trained user is to record scores. It is the task of the rater to assign a score for each item given information about an individual and in consultation with the scoring criteria. It may be useful to clarify this with an example from the Static-99R (Helmus, Thornton, et al., 2012), a tool designed to predict sexual reoffending. In rating the item named "Prior sex offences", the scoring manual instructs raters to give a score of 0 to men with no prior sex offence convictions, a score of 1 to men with one prior conviction, a score of 2 to men with two or three convictions, and so on. This process is known as coding. Scores for all items are then summed (but sometimes more advanced calculations are used) to give an overall risk score which corresponds with a probability of reoffending within a particular time frame (e.g., 5 years) and a risk "category" (e.g., Very Low Risk, Average Risk, Well above Average Risk etc.; Hanson et al., 2017). In short, Skeem and Monahan (2011) define actuarial tools as those that identify empirically validated risk factors, dictate a method for scoring these factors, establish a procedure for combining scores, and provide an estimate of risk.

The second generation of risk assessment was also defined by use of risk factors that are static in nature, meaning they cannot change over time (Bonta & Andrews, 2017). Research has identified a number of such examples related to sexual reoffending such as early onset of sexual offending and having prior convictions for sexually offending against

strangers (Hanson & Bussiere, 1998). Several studies have found actuarial static risk assessment tools possess superior predictive accuracy compared to UPJ (Andrews et. al, 2006; Hanson, 2009). Bonta and Andrews (2017) reported this a rare consensus among researchers in the field of forensic psychology. These tools are thought to have better predictive accuracy than UPJ for two main reasons. The first is that the inclusion of all risk factors is empirically justified (Bonta & Andrews, 2017). The second is that actuarial methods leave much less room for interpretation and therefore bias on the part of the assessor (Kemshall, 2001). Another benefit of using these tools is that the static (or historical) factors within them are relatively easy to assess because coding them often does not usually require lengthy interviews as much of the required information is typically already in a database. This in turn also makes using these scales cost effective (Department of Corrections, 2009). There are however notable downsides to actuarial static risk scales. First, because of their statistical underpinnings, it has been claimed they have no theoretical basis and largely ignore factors that are theoretically relevant, such as having pro-criminal associates (Bonta & Andrews, 2017). More importantly however, tools that rely heavily on static factors cannot detect change in risk over time. If an individual enters a psychological programme or through other methods achieves meaningful change, this will not be reflected in their score on a follow-up static risk assessment (Bakker et al., 1999). These tools also fail to identify risk factors that can be altered in order to lower risk (treatment targets) (Beggs & Grace, 2010). In theory, a clinician that can identify treatment targets can provide more individualised and therefore effective treatments (de Vries Robbé, de Vogel, Koster et al., 2015).

The third generation of risk assessment (Risk/Need scales) was defined by its inclusion of factors that are subject to change, labelled “dynamic risk factors” or “criminogenic needs” (Bonta & Andrews, 2017). Commonly cited examples in sexual offending literature are offence-supportive attitudes and sexual preoccupation (Mann et al.,

2010). Along with the already mentioned advantages of using dynamic risk factors, researchers are provided the opportunity to more easily evaluate the efficacy of psychological interventions. If a psychological intervention is shown to be associated with reductions in dynamic risk factors, this may indicate the intervention is effective while the treatment sample remains in prison. Empirically measuring treatment efficacy by releasing and following up on treated individuals over several years and comparing recidivism rates to those from untreated individuals remains a widely used and valid research paradigm (Bonta & Andrews, 2017), but the introduction of Risk/Needs scales provided an alternative (Papalia et al., 2020).

Somewhere on a spectrum between actuarial and unstructured methods of risk assessment lie Structured Professional Judgement (SPJ) tools. According to Skeem and Monahan (2011), SPJ tools are typified by the identification of risk factors and measurement of these, but do not dictate how the scores should be combined to estimate risk. Instead, a final risk estimation is made through structured decision making by trained experts (Douglas & Kropp, 2002). Webster et al., (1997, as cited in Skeem & Monahan, 2011) argue that professional judgement is important to integrate into risk assessment because it may be that a person is highly likely to reoffend due to the presence of only one risk factor, and this will not be reflected in scores on actuarial tools. However, it has been found that in the prediction of sexual recidivism, actuarial tools show higher accuracy than SPJ tools (Hanson & Morton-Bourgon, 2009). Murrie et al. (2009) suggested the less ambiguous nature of actuarial tools leave less room for interpretation and bias when compared with SPJ tools.

At this juncture, it may be useful for the reader to understand how risk assessment tools are evaluated. The accuracy of a tool in predicting recidivism (i.e., predictive validity) can be assessed through two types of experimental designs, prospective and retrospective (Singh et al., 2011). In prospective designs, the individuals are rated on a tool at the

beginning of a follow-up period (e.g., just before release from prison) and researchers track the individuals (usually over years) and record who reoffended by the end of this period. In retrospective designs, coders rate individuals using archived information from before the beginning of the follow-up period, but this coding occurs after the follow-up period has ended. To protect against bias, coders are blind to recidivism outcomes at the time of coding. Prospective designs have the advantage of enabling researchers to interview the individuals face to face to gain a more thorough understanding of those assessed, while in retrospective studies, researchers are limited to analysing information in case files. Retrospective designs have the advantage of giving researchers the opportunity to evaluate the predictive validity of a tool over a long follow-up period (e.g., over 10 years) without actually having to wait that long to collect and analyse recidivism data (Singh et al., 2011). Studies that utilise a retrospective design are usually conducted earlier in the tool development and validation process, with prospective studies usually being conducted after long-term predictive validity has been established (Olver et al., 2014). There are two types of analyses that can be conducted to measure the extent to which scores are related to recidivism outcomes (Bonta and Andrews, 2017). Originally, point biserial correlations were calculated between scores on a particular tool and recidivism outcomes (coded as 1 = reoffended, and 0 = did not reoffend). These analyses are no longer used as it was found that they are biased by base rates in reoffending. In more recent research, Receiver Operating Characteristic (ROC) analyses are used as they are unaffected by base rates (Rice & Harris, 1995). ROC analyses produce an Area Under the Curve statistic (AUC) which represents of the accuracy of an assessment tool in predicting a particular dichotomous outcome (e.g., recidivism). AUCs vary between 0 and 1.0, with 1.0 representing perfect prediction of the outcome variable and .5 representing chance discrimination (scores unrelated to the outcome variable; Bonta & Andrews, 2017). In predicting recidivism, AUC values can be interpreted as the probability that a randomly

selected recidivist will have a higher risk score than a randomly selected non-recidivist (Fawcett, 2006). In predicting recidivism, AUC values over .7 are considered moderate to large and values over .75 are considered large (Douglas et al., 2010, as cited in de Vries Robbé, de Vogel, Koster et al., 2015).

Researchers assessing risk tools are also interested in incremental validity. That is the degree to which a tool predicts a phenomenon of interest (e.g., recidivism) relative to other measures (Haynes & Lench, 2003). Incremental validity is assessed through hierarchical logistic or Cox regression analyses that allow researchers to first enter a controlled variable (e.g. another assessment tool) into a prediction model, and then the target tool is added to see if it improves the model and explains additional variance in the outcome variable (e.g. recidivism; Hunsley & Meyer, 2003). Finding incremental validity means that the tool is not simply measuring the same underlying constructs as other recognised assessment tools (or measures) but adding predictive power.

If a tool is shown to have poor predictive accuracy this can be for a number of reasons. One such threat to accuracy is poor interrater reliability. Interrater reliability is defined as the degree of agreement and correlation among two or more coders rating the same group of individuals (Koo & Li, 2016). Poor interrater reliability can arise from a range of differences between raters, such as inconsistencies in conscientiousness, the information being analysed to make judgements, their interpretations of this information, and/or differences in the interpretations of the scoring criteria. Establishing adequate interrater reliability is particularly important in the development of new measures and is a prerequisite for examining predictive validity (Field, 2013). Without establishing adequate interrater reliability, researchers do not have not a reliable construct from which to assess predictive validity (Seidlecki & Albert, 2017). The degree of agreement and correlation between the coders' ratings on items, tool domains, and overall scores are typically quantified by

intraclass correlation coefficients (ICCs). ICCs range between 0 and 1 with higher scores indicative of higher reliability, with scores between .75 and .9 indicative of good reliability and scores over .9 indicative of excellent reliability (Koo & Li, 2016).

Researchers evaluating psychometric measures (including risk assessment tools) are also interested in "construct validity". There has been controversy in scientific literature in regard to the definition of construct validity (Cording & Beggs Christofferson, 2017), but it is most commonly defined as "the degree to which a test measures what it claims, or purports to be measuring" (Brown, 1996, p. 231) or something similar (Caldwell et al., 1988). Using this definition, it follows that other forms of validity including predictive validity fall under the umbrella of construct validity, which was posited in Messick's (1989) widely cited paper. However, in forensic risk literature, predictive and construct validity are generally seen as separate entities, with construct validity being assessed by evaluating correlations between tools. Evaluating the construct validity of risk assessment tools involves scoring the same set of individuals on: the target tool; established and validated tools that measure related constructs; and tools or measures thought to be measuring unrelated constructs (Canales et al., 2009; Murphy & Davidshofer, 2004; Willis et al., 2020). Some forensic researchers divide construct validity into two forms. "Convergent validity" is established when scores from the tool being assessed are found to significantly correlate with scores from an established and validated measures of related constructs (e.g., static risk; Campbell & Fiske, 1959). "Discriminant validity" or "divergent validity" is established when a scores from the tool do not significantly correlate (or at all) with established measures of constructs theorised to be different. Risk assessment tools are also evaluated through assessing "concurrent validity" which is performed by analysing correlations between scores on the target tool and scores on other measures or tools that have been shown to be empirically associated with recidivism. This is done to make inferences about the target tool's ability to predict a real

world outcome (e.g., recidivism), rather than to evaluate if the tool is measuring the underlying construct it is claiming to (i.e., convergent and divergent validity; Field, 2013; Murphy & Davidshofer, 2004).

Commonly Used Sexual Risk Assessment Tools

In 2017, Kelley et al. (2020) surveyed 119 forensic professionals to investigate how they carry out sexual recidivism risk assessments with adult males. Eighty eight percent of responders were working in the United States of America, 7.2% were working in Canada, and the remaining 4.2% were working in various countries including the United Kingdom and Australia. Survey results indicated the most widely used sexual risk assessment tool was the Static-99R, a well-validated actuarial tool comprised of 10 static items. In a meta-analysis of 23 studies (n = 8,106), the tool was shown to yield moderate to large AUCs in the prediction of sexual recidivism (AUC = .71; Helmus, Hanson et al., 2012). Studies assessing the interrater reliability of Static-99R scores have yielded good ICCs ranging from .78 to .89 (Hanson et al., 2014; McGrath et al., 2012). According to Kelley et al. (2020), the second most used assessment tool among survey responders was the Stable-2007 (Hanson et al., 2012), an actuarial tool comprised of 13 "stable" dynamic risk factors, that is, factors capable of changing over months or years. In a meta-analysis of 12 studies (n = 6,845) the overall AUC in predicting sexual recidivism was .67 (Brankley et al., 2019). In the same publication, it was found that across four studies, the median ICC was .90, indicating excellent interrater reliability. Results from Kelley et al. (2020) showed the third most used tool was the Sexual Violence Risk-20 (SVR-20; Boer et al, 1997), an SPJ tool comprised of 20 items and measures both static and dynamic risk factors. Studies assessing the predictive validity of this tool in regard to sexual recidivism have yielded AUCs ranging from .72 to .83, with larger more robust studies yielding AUCs closer to the bottom of this range (Dempster, 1998, as cited in Jackson, 2016; de Vogel et al., 2004; Rettenberger et al., 2011). Hart and Boer (2009,

as cited in Rettenberger et al., 2009) analysed results from studies from Canada, Spain, Sweden, Austria, the Netherlands and Germany, and found overall, the interrater reliability of SVR-20 scores was at least fair, and in over half of these studies reliability coefficients could be classed as excellent. Finally, results from Kelley et al. (2020) indicated the fourth most used risk assessment tool among survey responders was the Violence Risk Scale: Sexual Offender Version (VRS-SO; Wong et al., 2003-2017). The VRS-SO is an actuarial tool comprised of 17 dynamic and 7 static risk factors predictive of sexual recidivism. Studies assessing the predictive validity of this tool in regard to sexual recidivism have yielded AUC values ranging between .61 and .80 (Beggs & Grace, 2010; Olver et al., 2007; Olver et al., 2014; Sowden, 2013; Sowden & Olver, 2017), but values have tended to fall around .70. Studies that have evaluated interrater reliability have generally reported good ICCs with values falling between .68 and .93 (Beggs & Grace, 2010, 2011, Eher et al., 2015; Olver et al., 2007; Sowden & Olver, 2017).

Protective Factors in Risk Assessment

It has been argued that modern risk assessment tools focus almost exclusively on factors that are associated with increased rates of recidivism, meaning factors that are associated with decreased rates of recidivism are overlooked (de Vries Robbé, Mann et al., 2015). By ignoring these protective factors, it is argued that important information is being missed, leaving risk assessment unbalanced (Rogers, 2000). There is debate in forensic psychology literature about the definition of protective factors (Cording & Christofferson, 2017). Some researchers interpret the concept as simply the absence of a risk factor (Zeng et al., 2015). Other researchers define them as factors that interact with risk factors to lessen their impact on the likelihood of future offending (Farrington et al., 2016). The current study utilises the definition put forward by Willis et al. (2020) which was influenced by the works

of de Vogel et al. (2012) and de Vries Robbé, Mann et al. (2015). Their definition relates to factors protective against sexual offending and is as follows:

> Those factors that are theoretically or empirically associated with reduced rates of sexual or violent recidivism in individuals with at least one apprehension for a sexual offence as an adult. They must signal the presence of a strength, not merely the absence of a risk factor or deficit. (Willis et al., 2020, p. 2)

They go on to clarify that some protective factors may represent the opposite end of a risk spectrum for a given factor (that has risk at the other end), and a person may have protective and risk factors present simultaneously on the same spectrum. Take for example the influence of peers: having antisocial associates (or negative social influences) is an established risk factor for sexual recidivism (Mann et al., 2010). Having prosocial associates has also been shown to be associated with reduced sexual recidivism which represents a protective factor (Willis & Grace, 2009; Willis et al., 2017-2019). It is of course possible to have both of these factors present, and it follows that both should be considered in a balanced risk assessment. Other protective factors are theorised to be completely independent of known risk factors, such as life goals and medication (de Vogel et al., 2012; Willis et al., 2020).

According to Willis et al., (2017-2019), theorised mechanisms through which protective factors lower risk of reoffending can be divided into two broad categories: control and prosocial reward. Control "refers to processes that restrain urges to engage in antisocial behaviour or, more broadly, mitigate the operation of risk factors" (p. 4). This control can be exerted from different sources including: internal sources (self-control); social/informal sources (being accountable to prosocial colleagues/friends/family members); or formal sources (imprisonment, parole conditions etc.; Thornton et al., 2017). Prosocial reward "refers to processes that lead the person to experience a prosocial life as rewarding" (Willis et al., 2020, p. 4). Whereas control mechanisms operate through the control and punishment of

antisocial living, prosocial reward mechanisms operate through the rewards inherent in prosocial living. For example, a person may feel a sense of mastery in having prosocial leisure activities, or a sense of relatedness in having a prosocial intimate relationship, and these prosocial rewards motivate prosocial behaviours and mitigate the likelihood of offending.

The theorised benefits of including protective factors in risk assessment can be loosely divided into two overlapping groups: the potential benefit to the accuracy of recidivism predictions, and the benefits to therapeutic interventions. Pertaining to predictive validity, it is posited that while current risk assessment tools can perform better than the unstructured methods of the past, there is still room for improvement. According to Cording and Christofferson (2017), risk dominant tools with the highest predictive validity tend to yield AUC values of around .70. While it has been theorised that these AUC values may be reaching a ceiling (e.g., due to issues surrounding attaining accurate recidivism data), it has been argued that attending to previously ignored protective factors could raise these values further (de Vries Robbé, de Vogel, Koster et al, 2015). It has been posited that using deficit focussed risk assessment leads to over-predictions in risk estimates as compensatory factors are not taken into account (de Vries Robbé, Mann et al., 2015; Rogers, 2000). Willis et al. (2020) argued that modern and well-established risk assessment tools that rely only on risk factors "are not able to identify groups with truly high recidivism rates" (p. 2) and cite the three following findings as evidence. In their independent retrospective validation study of the VRS-SO, Beggs and Grace (2010) found that only 56.2% of men who had VRS-SO scores placing them in the highest risk category reoffended after an average of 12.2 years. Using the same tool, Olver et al. (2007) found that 70% men in the highest risk category sexually reoffended after a period of 10 years. Similarly, Fernandez et al. (2014, as cited in Willis et al., 2020) investigated the predictive validity of combining Static-99R and Stable-

2007 scores, and found that most of the individuals placed in either high or very high risk categories were not charged or reconvicted of sexual offences over a 5 year follow-up period. In relation to these findings, Cording and Beggs Christofferson (2017) argued that "high risk" individuals do not appear to reoffend at a rate as high as one might expect. It should be noted however, that these somewhat low percentages may not only be affected by the ability of these tools to identify groups with truly high recidivism rates, but also by the cut-off scores used to determine who is "high risk", and the fact that a high proportion of sexual crime goes unreported to the police. Regardless, it remains a priority among researchers to discover why some individuals deemed high risk reoffend, and why some do not. It has been argued that what differentiates them may, to some extent, be the presence of protective factors (Willis et al., 2020). Studies assessing the predictive validity of tools that utilise protective factors are reviewed later in the introduction.

Over-estimating risk levels can have negative implications for the individuals themselves, taxpayer spending, and for treatment outcomes (Cording & Christofferson, 2007, Miller, 2006). As alluded to previously, if an individual's risk is overestimated (e.g., by an unbalanced assessment process), overly restrictive sanctions could be placed on the person which may violate their rights (Beggs, 2008), and potentially waste resources. Additionally, it is common practice for those deemed high risk to be placed in high intensity treatment programmes, and there is evidence to suggest that when a low risk individuals are placed in these programmes, their risk of reoffending is counterintuitively raised (Bonta et al., 2000; Lowenkamp & Latessa, 2004).

Beyond predictive accuracy and its implications, it is posited that integrating protective factors into forensic assessment can enhance therapeutic interventions and therefore improve therapeutic outcomes (de Vries Robbé & Willis, 2017). As noted by Cording and Beggs Christofferson (2017), two of the most influential models of criminal

rehabilitation support incorporating strengths into assessment and therapy. The Good Lives Model (GLM; Ward & Stewart, 2003) in particular utilises a strength-based approach. The theory dictates that all humans are naturally goal directed and want to seek out a combination of 11 listed "primary human goods", which are intrinsically beneficial actions, outcomes or states of affairs. This list includes common goals such as inner peace, excellence in work, pleasure, and creativity. The theory stipulates that people vary in the number and combination of goods they value, and vary in the strategies they use and capacities they have to acquire them. Offending is said to occur when there are flaws in the number and/or combination of goods sought, and/or the strategies and capacities relevant to attainment. It is the role of the therapist to identify these flaws, and help the client formulate and implement a plan that facilitates the attainment of goods via fulfilling and prosocial means. Of particular relevance to protective factors, part of this process involves identifying, enhancing and creating personal strengths. Similarly, the highly influential Risk Needs Responsivity (RNR; Bonta and Andrews, 2017) model of rehabilitation also supports the assessment of strengths (Cording & Beggs Christofferson, 2017). The "Responsivity" principle dictates that treatment style and mode should be tailored to the specific characteristics of the individuals being treated, including their strengths. Further, the model posits that strengths should be assessed to enhance prediction and treatment responsivity.

It has been suggested that shifting focus towards strength-based assessment could improve the rapport between clients and clinicians (Cording & Beggs Christofferson, 2017, de Vogel et al., 2012; de Vries Robbé, Mann et al., 2015, de Vries Robbé & Willis, 2017). Attrill and Liell (2007, as cited in de Vries Robbé & Willis, 2017), interviewed men in prison and discovered they thought the deficit focused nature of risk assessment led to an overly negative view of their risk level, and yielded biased information that was used to justify unfair risk management decisions. De Vries Robbé & Willis (2017) argued that this

negativity towards risk assessment among individuals serving sentences of imprisonment may manifest in poor rapport between clients and clinicians, which could hinder treatment engagement and motivation on the part of the client. Therefore, they suggested that by incorporating protective factors, this could improve client engagement and motivation, leading to better treatment outcomes. This point was echoed by de Vogel et al. (2012) who noted that taking a positive and collaborative approach to assessment can be motivating for both clinicians and their clients.

The SAPROF-SO and SAPROF

The SAPROF-SO is a new tool that was designed to assess hypothesised protective factors against sexual recidivism and was developed with the aim of informing treatment and measuring change, as well as better informing reoffending predictions (Willis et al., 2017-2019). The tool contains 23 dynamic and one static item, and is intended for use with biologically born males over 18 years old. From an extensive search, the tool appears to be the first assessment tool that both exclusively assesses protective factors against sexual offending, and was developed for use with adults. Another tool, The Desistence for Adolescents who Sexually Harm (DASH-13; Worling 2013) was developed to assess protective factors against sexual recidivism in adolescents, but empirical research validating this tool is currently lacking. The SAPROF-SO is currently in a pilot stage and is in development towards being an actuarial tool (Willis et al., 2020). In line with its philosophy of bringing balance to risk assessment, the SAPROF-SO is intended to be used alongside risk measures. From results of a yet to be finalised factor analysis, the items within the SAPROF-SO are divided into two overarching domains (Thornton & Kelley, 2020). The Personal domain includes 19 items that relate to: internal protective capacities; having a prosocial and/or adaptive identity; the extent to which an individual is connected to prosocial others and/or activities; and/or having housing and financial stability. The Professionally Provided

Support domain consists of five items that pertain to support provided from external sources. Descriptions of SAPROF-SO items and the two domains can be found in Table A1 (Appendix A). This table was adapted from the supplementary materials from Willis et al. (2020).

The primary aim of the current study is to evaluate the validity of the SAPROF-SO in the prediction of sexual reoffending, and is the first to do so. What follows is a brief description of the tool from which the SAPROF-SO was adapted, the Structured Assessment of Protective Factors for violence risk (SAPROF; de Vogel et al., 2012), and a review of studies that have assessed predictive validity of the SAPROF. Following this review, the development of the SAPROF-SO is described and preliminary research assessing construct validity (convergent and divergent), and the interrater reliability of SAPROF-SO scores (Willis et al., 2020) is discussed.

The SAPROF-SO is an adaptation and extension of the SAPROF, an SPJ tool developed for the assessment of protective factors for general violence risk, but "intended to be applicable both to violent and to sexually violent offenders" (de Vries Robbé, de Vogel, Koster et al., 2015, p. 53). The SAPROF contains two static and 15 dynamic items.

Studies assessing the predictive validity of the SAPROF have generally yielded favourable findings, particularly in the prediction of violent recidivism. De Vries Robbé et al. (2011) investigated the accuracy of the SAPROF in the prediction of violent reoffending. In their retrospective study, the SAPROF was scored from case files of 105 male psychiatric patients who had a history of violence and were housed in a secure treatment facility. Because the SAPROF is an SPJ tool, final scores are influenced by both items scores and clinical judgement (Final Risk Judgement scores). In addition to calculating scores this way, the authors also simply summed the item scores to give a SAPROF Total score, which is in line with actuarial methods, and is how the SAPROF-SO is scored (Willis et al., 2017-2019).

This enabled researchers to compare the predictive accuracy of the two score calculation methods. It was found that SAPROF Total scores were significantly predictive of violent offending at follow-up times of one year (AUC = .85, $p < .01$), two years (AUC = .80, $p < .01$), and three years (AUC = .74, $p < .01$). Further, AUCs for SAPROF Total scores were higher than those for Final Risk Judgement sores, which Willis et al. (2020) noted is somewhat consistent with research suggesting actuarial tools outperform SPJ tools in the assessment of sexual recidivism risk (Hanson & Morton-Bourgon, 2009). Overall, the results from this study support the claim that empirically derived protective factors can be used to predict recidivism.

De Vries Robbé, de Vogel, Koster et al. (2015) retrospectively investigated the accuracy of the SAPROF in the prediction of violent and sexual reoffending, using a sample of 83 men with convictions for "hands on" sexual offending. They found that SAPROF Total scores were significantly predictive of sexual offending at both follow-up times of 3 years (AUC = .76, $p < .05$) and 15 years (AUC = .71, $p < .01$). SAPROF Total scores were also significantly predictive of violent offending and AUCs from theses analyses were slightly higher for both follow-up times, than for sexual reoffending. This makes sense given the SAPROF has an increased focus on factors predictive of violence rather than sexual offending. The researchers also found that SAPROF Total scores remained significantly predictive of sexual recidivism at both follow-up times, after simultaneously controlling for SVR-20 scores and scores on an SPJ tool designed to assess risk of violence – the Historical Clinical Risk Management-20 (HCR-20; Webster et al., 1997). These regression results support the claim that the accuracy of risk assessments could be improved by integrating protective factors. Further, neither HCR-20 nor SVR-20 scores were shown to significantly predict sexual recidivism at either follow-up time. The finding that the SAPROF outperformed the well-validated and widely used SVR-20 was particularly promising given

the SAPROF has an increased focus on violence risk and the SVR-20 was designed to assess risk of sexual offending.

Yoon et al., (2018) retrospectively investigated the accuracy of the SAPROF in the prediction of general, sexual, nonsexual violent, and sexually violent recidivism, using a sample of 450 men convicted and imprisoned for sexual offences. They found SAPROF Total scores were not predictive of sexual reoffending (AUC = .53) but were significantly predictive of general recidivism (AUC = .62, $p < .01$), nonsexual violent recidivism (AUC = .63, $p < .01$), and sexually violent recidivism (AUC = .61, $p < .05$), after an average follow-up time of 5.8 years post release. These results support the earlier findings that the SAPROF is better at predicting violent than sexual offences. As noted by Willis et al. (2020), the small AUCs yielded in this study may be explained by a methodological difference. SAPROF scores were coded from reports written at prison intake, and AUCs may have been higher if they were coded from discharge plans. This is because many SAPROF items are dynamic and therefore may change during the course of a prison sentence (e.g. because of treatment), and additionally, "release environments differ in the extent to which they afford opportunities for the expression of different protective factors which intake assessments cannot account for" (Willis et al., 2020, p. 3).

Turner et al. (2016) retrospectively investigated the accuracy of the SAPROF in the prediction of violent, sexual, and general reoffending, using a sample of 277 men convicted and imprisoned for sexual offences against children. They found SAPROF Total scores were not significantly predictive of either violent offending (AUC = .60) nor sexual offending (AUC = .52), but were significantly predictive of general offending (AUC = 0.63, $p < .05$) after an average follow-up time of 5.71 years post release. The authors reasoned that small AUCs were found in their study because their sample consisted of people convicted of child sexual offences. They suggested that the SAPROF items Intelligence and Social Network

may act as risk factors rather than protective factors for men who have committed this type of offence, especially for those that committed their crimes while working with children. They reported this may be because those that are well connected to others and appear intelligent may have greater access to children. However, these suggestions are not supported by child sexual offending research investigating intelligence (Beggs & Grace, 2008; Willis & Grace, 2008), and social influences (Willis & Grace, 2009). Willis et al., (2020) also noted that in this study, SAPROF scores were coded at prison intake, which likely deflated AUCs for the aforementioned reasons.

SAPROF-SO: Development and Preliminary Research

According to Willis et al. (2020) the SAPROF-SO authors originally perceived there was significant overlap in protective factors relevant to violent and sexual offending, and therefore intended to create a supplementary tool to be used alongside the SAPROF with some changes in items and refinements to criteria. However, upon reviewing empirical research, the authors decided that the magnitude of required changes necessitated the creation of a new stand-alone tool. The changes were not only inspired by the increased focus on sexual offending in the SAPROF-SO, some changes were made to convert it to being actuarial, and improve the general utility of the tool. To improve utility, the response scale for each item was expanded from 3 to 5 points. This was done to capture greater variance in protection and better detect changes. In an effort to develop an actuarial tool and increase interrater reliability, scoring instructions were made more explicit and specific (Willis et al., 2020). To facilitate clinical reasoning, item descriptions were expanded to include theorised underlying mechanisms through which each protective factor lowers risk (Willis et al., 2020). The SAPROF-SO has many items that are environmental in nature (Housing Stability, External Control, Supervised Living etc.). Therefore, for each item, users are able to give scores for the client's current context (e.g. prison) and if a change in living situation is

imminent (e.g. release from prison) the coder can also give scores for a future context (e.g. post release living situation; Willis et al., 2017-2019). Many items from the SAPROF were retained in the SAPROF-SO, and through reviewing theoretical and empirical literature exploring factors related to both sexual and general offending, three items were significantly refined, and seven were added. Since the conception of the SAPROF-SO, further changes were made to improve the tool following consultation with experts and testing it on cases (Willis et al., 2020). These changes included removing one item and renaming two others. Finally, the first iteration of the SAPROF-SO contained five theoretically informed domains (one of which was the Professionally Provided Support domain), but preliminary results of a yet to be finalised factor analysis (Thornton & Kelley, 2020), suggested two overarching domains which were used in the current study. The process of combining scores to allocate individuals into risk categories and estimate overall protection is currently under development (G. M. Willis, personal communication, December 22, 2020).

Willis et al. (2020) evaluated the interrater reliability of SAPROF-SO scores and assessed construct validity (convergent and divergent) using both a high risk sample and a routine sample. The high risk sample consisted of 40 men who had been convicted of sexually violent offences and had a psychological disorder that increased their risk. These men were participating in a psychological treatment programme at a secure facility in the United States (Sand Ridge Secure Treatment Center), and SAPROF-SO scores were informed by analysing treatment reports. The average Static-99R score for the high risk sample was 5.52 which falls in the Above Average risk category (Hanson et al., 2017). The routine sample consisted of 40 men in New Zealand who were previously imprisoned for index sex offence convictions and at the time of coding were released. The average Static-99R score for the routine sample was 2.38 which falls in the Average risk category. For the routine sample, SAPROF-SO scores were informed by review of: an audio recording of a telephone interview

with each participant conducted by a person trained in using the tool; psychological reports and reports completed by probation officers. For both samples, only Current context scores were calculated.

The interrater reliability of SAPROF-SO Total scores was excellent for both samples. The ICCs for the high risk and routine samples were .90 and .94, respectively. The ICCs for the Professionally Provided Support domain were good for both samples with ICCs of .85 for the high risk sample and .87 for the routine sample. Willis et al. (2020) used the earlier version of the SAPROF-SO that divided items into five domains, therefore the results for other domains are not reported here. For each item, the percentage of cases in which coders were within 1 point of agreement (on the five point scale) was also high, with only two items (Social Network and Emotional Connection to Adults) in the high risk sample failing to meet at least 70% agreement. Across all items, the average percent agreement within 1 point was 84.3 for the high risk sample and 86.5 for the routine sample.

Willis et al. (2020) also found support for construct validity. For the high risk sample, the degree to which SAPROF-SO scores correlated with both Static-99R and VRS-SO scores was measured. For the routine sample, the degree to which SAPROF-SO scores correlated with scores from both the Static-99R and the Dynamic Risk Assessment Scale for Offender Re-Entry (DRAOR; Serin, 2007) was measured. The DRAOR is a SPJ tool that was designed to assess general risk of recidivism in men on probation, and is moderately accurate in doing so (Hanby, 2013). Further, the DRAOR contains a domain that consists of protective items (Protective Factors). Supportive of divergent validity, the Willis et al. (2020) found SAPROF-SO Total scores did not significantly correlate with Static-99R scores from either sample. This finding was predicted and supports divergent validity because the SAPROF-SO is a measure of dynamic protective factors rather than static risk factors. Divergent validity was further supported by the finding that SAPROF-SO Total scores did not significantly

correlate with VRS-SO Pretreatment Dynamic Risk scores, a scale that captures historical functioning (as opposed to ongoing functioning which is captured within the SAPROF-SO and other VRS-SO scales). In support of convergent validity, it was found SAPROF-SO Total scores were significantly positively correlated with changes in dynamic VRS-SO scores, indicating that reductions in dynamic risk factors were associated with a greater presence of protective factors. The significant correlations between the SAPROF-SO and change in the VRS-SO domains: Sexual Deviance; and Treatment Responsivity; were particularly important findings because these domains capture risk factors specific to sexual offending. These findings indicate that the SAPROF-SO is likely capturing protective factors specific to sexual offending. Further confirming convergent validity, SAPROF-SO Total scores were significantly positively correlated with DRAOR Protective Factor scores, and significantly negatively correlated with all DRAOR scales indicative of ongoing general risk. Across all analyses, Willis et al. (2020) noted that the strongest correlations were around .7, and the SAPROF-SO was shown to be reliable. Therefore, the authors concluded "the SAPROF-SO correlates with measures of risk in a way consistent with its construct validity, but it clearly also captures substantial variance that is independent of the risk measures" (Willis et al., 2020, p. 16). The authors identified that a limitation to this study was that coders were not always blind to risk scores, and in some cases narrative descriptions of risk assessment scores were used to inform coding. This would theoretically inflate correlations in the convergent validity analyses.

The Current Study

The current study is the first to assess the accuracy of the SAPROF-SO in the prediction of recidivism. As mentioned, the primary aim of the current study was to evaluate the validity of this tool in the prediction of sexual reoffending, but validity in the prediction of general and violent recidivism was also evaluated. The current study additionally

investigated if the tool can provide incremental validity over and above a well-validated measure of static risk (Static-99R). To aid in ruling out poor reliability as a threat to accuracy, interrater reliability was calculated. The current study represents a larger effort to develop and validate the SAPROF-SO, and a small step towards the tool meeting the requirements to be deemed actuarial and being utilised in forensic practice. Finding predictive and incremental validity would support the claim that integrating the SAPROF-SO (and protective factors by extension) into risk assessment could improve their accuracy.

The present study utilised a retrospective design and SAPROF-SO scores were informed by analysing the treatment reports and related documents of 210 men who participated in a prison-based treatment programme for men who have sexually offended against children (Kia Marama, Rolleston Prison, New Zealand), between 1993 and 2000. The current study utilised recidivism data previously used in Beggs & Grace (2010) and the average follow-up time was 12.24 years. Before coding began, a Retrospective Scoring Guide was developed by the author (Appendix B). This guide was developed to maximise interrater reliability, aid coders in finding pertinent information, and to minimise item omissions. Because the SAPROF-SO was developed for clinical use and information pertaining to protective factors during the period participants completed treatment was scarce, it was not ideal for use in a retrospective study. The guide minimised omissions by allowing coders to use information beyond that recommended in the manual to make decisions regarding the presence of protective factors. To ensure these changes were congruent with the theoretical underpinnings of the SAPROF-SO, the guide was critiqued, amended, and approved by an author of the SAPROF-SO.

It was hypothesised the SAPROF-SO scores would be significantly predictive of sexual offending and to a higher extent than general and violent offending. This hypothesis was based on the fact that items were empirically and theoretically derived, SAPROF Total

scores have been shown to be predictive of violent and sexual offending (and more predictive of violent than sexual offending; de Vries Robbé, de Vogel, Koster et al., 2015), results pertaining to construct validity (Willis et al., 2020), and because cases were coded from discharge plans. It was further predicted that Future context ratings would show superior predictive validity to Current context ratings, as many items are environmental in nature, and Future context ratings may reflect the environment in which the recidivism occurred. Further, it was hypothesised that SAPROF-SO scores would provide incremental validity over and above Static-99R scores. This prediction was based on the finding that SAPROF-SO scores did not significantly correlate with Static-99R scores in a preliminary study (Willis et al., 2020), and because the SAPROF-SO measures factors that are almost exclusively dynamic and reflect ongoing rather than historical functioning. Based on the interrater reliability findings from Willis et al. (2020), it was expected the SAPROF-SO would show excellent interrater reliability for Total scores and good to excellent interrater reliability for the Professionally Provided Support domain scores (for both Current and Future contexts).

Method

Participants

Case files were requested of 218 adult males who had participated in a treatment programme for men who have sexually offended against children while imprisoned at the Kia Marama Special Treatment Unit in Rolleston, New Zealand, between 1993 and 2000. These were the same case files as those used in previous research undertaken by Beggs and Grace (2010, 2011). Of this sample, four case files were in use and unavailable at the time of data collection, and four case files were irretrievable following the 2011 Christchurch earthquake. Therefore, the remaining 210 files constituted the total number used in the current study. At the time of entering the programme, these men were aged between 18 and 74, with an average age of 41.2 years (*SD* = 12 years). Ethnically, 77.6% were of Pākehā (New Zealand

European) descent, 20% were of New Zealand Māori descent, and the remaining 2.4% were from other ethnicities, including those from the Pacific Islands. When the participants entered the programme, they consented to their information being used for future research purposes.

Measures and Data Collection Procedures

SAPROF-SO – Pilot Version

The SAPROF-SO (Willis et al., 2017-2019) is a 24-item clinician-rated risk assessment tool that uniquely and exclusively measures protective factors to evaluate risk of sexual offending in adult men (Willis et al., 2020). The full list of items is presented in Table 1 and descriptions of SAPROF-SO items and the two domains are presented in Table A1 (Appendix A). In the current study, because the nature of the therapeutic alliance was not routinely documented, it was decided that this item (Therapeutic Alliance) would be omitted for all cases.

Using all available information, whether it be from behavioural observation, interview information or from reviewing case files, for each item, the SAPROF-SO scoring manual instructs raters to assign the individual the score that is most suitable given the stated scoring criteria. The manual stipulates that the coder should only consider information from within a specific time window, which varies from item to item. Typically, only information that pertains to the previous 6 to 12 months is analysed, but more historical information can be considered when coding some items that are thought to be more stable like prosocial sexual interests or adaptive schemas (Willis et al., 2020). Higher scores are indicative of stronger protection against sexual recidivism. Across all items, a score of 4 is assigned when the protective factor is “clearly present”, a score of 2 when it is “somewhat present” and a score of 0 when it is “not present”. These are referred to as anchor points. In the event the rater cannot decide between two anchor points, they are to assign a middle score of 1 or 3. If a rater cannot decide between a middle score and an anchor point, the anchor point is recorded.

Four items (Motivation for Managing Risk, Sexual Offence-Specific Treatment, Medication, and Therapeutic Alliance) have "not applicable" (N/A) or "not required" (N/R) response options along with numerical coding options.

The scoring manual also requires that for each item, the rater gives a score for their Current context (e.g., prison), and allows for a separate score to be assigned for a Future context if it is expected to change in the next 6 months (e.g., release from prison). The first 12 items of the SAPROF-SO measure factors that are estimated to be reasonably stable across time and contexts. Therefore, Current scores for these items are usually used as Future context scores (unless the rater has reason to believe these may change). Future context scores are coded separately for the remaining 12 items (Willis et al., 2017-2019). In the present study, Future context scores were recorded as the same as Current scores for the first 12 items and Medication, but were coded separately for the remaining 10 items (Therapeutic Alliance was omitted). Because the treatment programme is typically only available to individuals who are due to be released following treatment participation, Current context scores were informed by the prison environment and Future context ratings were informed by case information relating to post release plans. One participant was convicted and sentenced to further prison time while participating in treatment, therefore, his Future context scores were informed by information pertaining to the prison he was being transferred to. In calculating Total and domain scores for each context, item scores were summed.

Retrospective Scoring Guide

Because the SAPROF-SO was developed for clinical use and protective factors, being a newer concept to emerge in the field (Rogers, 2000), were not as commonly explicitly referred to in treatment documents as they are in more recent psychological practice, the SAPROF-SO was not ideal for use in a retrospective study. A further issue was that it was often a lengthy exercise to find specific information to code items amongst the complete case

file information. Therefore, before coding began, a supplementary scoring guide was developed for the purpose of this study, which instructed coders how to rate items given the unique and limited material available, and aided coders in finding specific information within files and documents (Appendix B). Additionally, it was developed to maximise interrater reliability. This Retrospective Scoring Guide was developed by the author of the current study by analysing over 20 case files from the current sample to assess the layout and scope of available information. To ensure that the Retrospective Scoring Guide was theoretically aligned with the SAPROF-SO, the content of was critiqued, amended and approved by an author of the SAPROF-SO.

According to the scoring manual, in circumstances when there is insufficient information, items are to be omitted. To minimise omissions, for some items, the Retrospective Scoring Guide instructed raters to consider information that is not usually analysed when coding the SAPROF-SO. When necessary, raters were to consider related information that predates the time window that is stipulated in the manual (but with more weight given to recent functioning). For example, when coding the Empathy item, coders were to consider evidence of empathic behaviours that took place before the stipulated time window (6 months), such as historical involvement in charity groups or caring for a vulnerable family member. For the Self-Control item, coders were to gauge restraint by not only searching for evidence of self-control over the previous 12 months, but by also examining older information pertaining to educational achievement, violent outbursts, drug and alcohol use and employment stability. In scoring Attitude Towards Rules and Regulations, history of non-sexual offences was considered, as was instability of employment history (which may indicate issues with rules and authority figures). History of having prosocial intimate relationships with adults was additionally examined when coding the Prosocial Sexual Interest item. Further, specific psychometric scores on file were also

analysed when coding Adaptive Schema, Empathy, Coping, Self-Control, Goal-directed Living, Attitudes Towards Rules and Regulations, and Financial Management. The details of which can be found in Appendix B.

For two items, the Retrospective Scoring Guide instructed the coder to deviate from the scoring criteria of the official manual. For the Self-Control item, scores were assigned based on the coder's own definition of self-control. Those judged to have "good" self-control were assigned a 4, those with "average" self-control were assigned a 2, and those with "poor" self-control were assigned a 0. This change was made because there was often scarce information pertaining to the efficacy and implementation of "self-control strategies". For the Sexual Self-Regulation Item, the coder was to slightly deviate from scoring clarifications in the manual and code each of the four listed "elements" separately as either present (1) or not present (0). The four elements relate to: (i) having a lifestyle devoid of opportunities to sexually offend; (ii) having strategies to negotiate high risk situations; (iii) having low levels of offence-related sexual impulses and not acting on those impulses; and (iv) having prosocial sexual thoughts and fantasies that are acted upon in a contextually-appropriate way, while also not using sex to cope with negative affect. The coder was to then calculate a total Sexual Self-Regulation score by giving those with zero or one elements present a zero, those with two elements present a one, those with three elements present a two, if they had all four elements present for over 12 months, they would receive a four, and otherwise, they would receive a three. The implications of using the Retrospective Scoring Guide will be presented in the discussion section of this paper.

SAPROF-SO Coding Procedure

Training was provided by the SAPROF-SO authors via video conferencing software to the author, a postgraduate clinical psychology student. Coding was informed by a comprehensive review of relevant material including each participant's final psychological

treatment report, and other documents dated before this such as court documents, treatment exercises and psychometric test results. The final treatment reports were written at the end of treatment and usually shortly before release from prison. The full list of documents that were examined can be found in Appendix B.

At the beginning of coding, 10.5% of cases (n = 22) were randomly selected and rated independently by both the author and a SAPROF-SO author and these scores were used to calculate the interrater reliability. These cases were coded by both raters in batches of two to three, and between batches discrepancies between item scores were discussed and resolved by consensus scoring. This process of discussion and consensus scoring provided the author with the opportunity to resolve any misunderstandings and allowed for further clarification of the scoring criteria. Additionally, this led to some minor changes being made to the Retrospective Scoring Guide. The remaining 89.5% of cases (n = 188) were then coded by the author alone. The cases that were scored by consensus and the cases that were rated by the student constitute the full data set that was used for the remainder of the statistical analyses (i.e., excluding the interrater reliability analyses). Both coders were blind to all recidivism outcomes at the time of coding.

Static-99R

The Static-99R is the most widely used and researched risk assessment tool that evaluates risk of sexual offending (Kelley et al., 2020; Reeves et al., 2018). This actuarial tool consists of 10 static risk items, and total scores range between -3 and 12 with higher scores indicative of higher risk. The most recent iteration of this tool has five risk categories, and has updated category labels designed to improve score interpretations in forensic and correctional settings (Hanson et al., 2017). Individuals scored between -3 and -2 are rated as “Very Low Risk”, scores of -1 to 0 indicate “Below Average Risk”, 1 to 3 indicate “Average Risk”, 4 to 5 indicate “Above Average Risk” and scores over 6 indicate “Well Above Average Risk”.

Static-99R scores for the current sample were calculated by transforming Static-99 scores that were used in previous research undertaken by Beggs and Grace (2010). The difference between these two versions are the coding rules for Age at Release from Index Sex Offence item. It was found that risk of sexually reoffending greatly decreases as men enter middle-age, and the scoring of this item was changed as a result (Helmus et al., 2009). This updated scoring method was able to be retrospectively applied to scores for the current sample. Both raters were blind to Static-99 and Static-99R scores at the time of coding the SAPROF-SO. As stated, across 23 studies (n = 8,106) the Static-99R was shown to yield moderate to large AUCs in the prediction of sexual recidivism (AUC = .71; Helmus, Hanson et al., 2012). Studies assessing interrater reliability have yielded ICCs indicative of good reliability (McGrath et al., 2012; Hanson et al., 2014).

Recidivism

Recidivism data used in the current study were extracted from the dataset utilised by Beggs and Grace (2010). This criminal history information was obtained from the computer database maintained by the New Zealand Department of Corrections with the final follow-up date being July 1, 2008. Beggs and Grace (2010) recorded whether each participant received convictions for sexual, violent, or general offences between release and the end of the follow-up period. As reported in their study:

> Sexual recidivism was defined according to the Static-99 scoring criteria for Category "A" offenses (Harris, Phenix, Hanson, & Thornton, 2003), that is, an offense with an identifiable victim (e.g., incest, sexual assault, exhibitionism). Category "B" offenses (i.e., no identifiable victim) were excluded, except for possession of child pornography. Violent recidivism was recorded when the offender had been convicted for a non-sexual offense against a person (e.g., assault, robbery, kidnapping). General recidivism was recorded for offenses that

> were neither sexual nor violent (e.g., possession of cannabis). (Beggs & Grace, 2010, p. 239)

For each individual in the current sample, the time at large prior to each reconviction, or to the end of the follow-up period was calculated. The average follow-up time was 12.24 years (*SD* = 1.86, range = 7.92 - 14.88 years).

Data Analyses

ICCs were calculated to assess the interrater reliability of SAPROF-SO scores. ROC analyses were conducted to assess the validity of the SAPROF-SO and Static-99R in the prediction of sexual, violent, and general recidivism. Cox regression analyses were run to evaluate whether the SAPROF-SO contributes incremental validity over and above Static-99R scores in the prediction of sexual recidivism. Statistical analyses were conducted with SPSS (Version 25.0).

Results

Descriptive Statistics

The number of valid ratings, along with the means and standard deviations for item, domain and Total scores are presented in Table 1. Future context statistics for the first 12 items are not presented as these are the same as current context statistics. In terms of item omissions, only three participants (1.4%) were missing more than three Current context scores and seven participants (3.3%) were missing more than three Future context scores. A substantial proportion of participants had at least one item score that was omitted. Fifteen participants (7.1%) had zero omitted Current context ratings and 72 participants (34.29%) had zero omitted Future context ratings. Overall, the mean number of omitted items per participant was 1.57 (*SD* = 0.83) for Current context ratings (range = 0-4), and 1.06 (*SD* = 1.35) for Future context ratings (range = 0-9). The most frequently omitted item was Leisure Activities which could only be scored for 35 participants (16.6%) for the Current context and

170 cases (81%) for the Future context. In Table 1, Sexual Self-Regulation scores were deemed "valid" when zero of the four "elements" required were omitted, which was evident in a relatively low proportion of cases (51%) for which ratings were obtained. In calculating domain and Total scores for each participant, omitted item scores were replaced with the mean domain score for that participant. Consistent with Willis et al (2020) "not applicable" (N/A) scores for Motivation for Managing Risk and Medication and were replaced with 0, indicating that protection was not present. No participants were assigned N/R for the Sexual Offence-Specific Treatment item. For the Sexual Self-Regulation item, omitted "elements" used in the calculation of total Sexual Self-Regulation scores were replaced with zero.

Across all participants, the average Total score was 39.08 (*SD* = 8.94) for the Current context and 34.38 (*SD* = 10.38) for the Future context. For many items, there was very little variability between participants in scores. For Current context ratings, all participants were assigned scores of 4 for Housing Stability, Supervised Living, External Control, and Sexual Offence-Specific Treatment because they were incarcerated and participating in the Kia Marama programme. For Future context ratings, 90.5% of participants were assigned a score of 4 for Sexual Offence-Specific Treatment. Additionally, 97.1% of participants were due to be released on parole and were subject to parole conditions, and therefore, these participants all received a score of 2 for External Control for the Future context. There was also little variability in scores for the Supervised Living item for the Future context, with only two participants being assigned scores higher than 2. Of the current sample, 94.3% were assigned N/A scores for the Medication item and therefore, these participants were all assigned a score of 0 in the calculation of Total and domain scores.

The mean Static-99R score for the current sample was 1.70 (*SD* = 2.46), which falls in the Average risk category.

Table 1

SAPROF-SO Descriptive Statistics

SAPROF-SO items and domains	Current context			Future context		
	n valid ratings	Mean	*SD*	*n* valid ratings	Mean	*SD*
Personal						
1. Intact cognitive functioning	210	3.25	1.19			
2. Secure attachment in childhood	201	1.44	1.38			
3. Adaptive schema	210	0.90	0.88			
4. Empathy	208	0.94	0.92			
5. Coping	210	1.25	0.95			
6. Self-control	210	1.44	1.17			
7. Sexual self-regulation	107	0.61	1.04			
8. Prosocial sexual interests	210	1.02	0.93			
9. Prosocial sexual identity	210	0.45	1.25			
10. Goal-directed living	207	0.70	0.87			
11. Motivation for managing risk	210	2.06	0.75			
12. Attitudes towards rules and regulations	210	1.85	1.16			
13. Work	206	0.35	0.77	205	0.64	1.02
14. Leisure activities	35	1.66	0.98	170	1.38	1.17
15. Social network	208	2.13	1.01	207	2.29	1.01
16. Emotional connection to adults	179	0.50	0.83	179	0.70	0.97
17. Intimate relationship	210	0.52	0.85	208	0.75	1.31
18. Housing stability	210	4.00	0.00	203	3.26	1.12
19. Financial management	209	2.01	0.83	206	2.18	1.09
Subtotal (possible range: 0 – 76)	210	26.91 (obtained range: 8.44 - 55.81)	8.99	210	27.00 (obtained range: 6.00 – 55.94)	10.10
Professionally provided support						
20. Sexual offence-specific treatment	210	4.00	0.00	206	3.79	0.78
21. Medication	210	0.18	0.78	210	0.18	0.78
22. Therapeutic alliance	[a]	[a]	[a]	[a]	[a]	[a]
23. Supervised living	210	4.00	0.00	203	1.48	0.76
24. External control	210	4.00	0.00	208	2.00	0.25
Subtotal (possible range: 0 – 16)	210	12.18 (obtained range: 12 – 16)	0.78	210	7.38 (obtained range: 0 – 12)	1.62
Total score (possible range: 0 – 92)	210	39.08 (obtained range: 20.44 – 67.81)	8.94	210	34.38 (obtained range: 11.40 – 63.94)	10.38

Note. SAPROF-SO = Structured Assessment of PROtective Factors – Sexual Offence version (Willis et al., 2017-2019). SAPROF-SO items are rated 0 – 4. Missing scores were replaced with case mean domain scores in calculation of subtotal and total scores. Not applicable ratings were replaced with scores of 0. For Sexual self-regulation scores, cases were deemed valid when zero elements were omitted.

[a] Therapeutic alliance was not coded in the current study due to a lack of relevant information.

Interrater Reliability

To aid in ruling out poor reliability as a threat to predictive accuracy, interrater reliability of SAPROF-SO scores was calculated. Using the scores of the 22 participants that were coded by two raters, single-rater ICCs (two-way mixed, absolute agreement) were calculated. Decisions regarding the types of ICC analyses used and the interpretations of these statistics (e.g., poor, moderate, good, or excellent) were informed by the widely cited paper by Koo and Li (2016). The single-rater ICCs for the Total Current and Total Future scores were .98, 95% CI [.93, .99], and .98, 95% CI [.95, .99], respectively, indicating excellent interrater reliability. These statistics were higher than what was reported in Willis et al., 2020, although their results also fell in the excellent range. More detailed results of interrater reliability analyses for items, domains and Totals across both contexts are available in Tables C1 and C2 (Appendix C).

The single-rater ICCs for the Personal domain were .98, 95% CI [.93, .99] for Current context ratings and .98, CI [.95, .99] for Future context ratings, indicating excellent interrater reliability. Comparisons to Personal domain results from Willis et al. (2020) cannot be made as they utilised an earlier version of the SAPROF-SO that divided the first 19 items into domains differently. It should be noted that there were zero valid cases (cases in which both raters were able to code the item) for Leisure Activities for the Current context. Additionally, there were only three valid cases for Sexual Self-Regulation. Missing scores were replaced with case mean domain scores in the calculations of ICCs for domain and Total scores. The ICC for the Professionally Provided Support domain for the Current context was not able to be calculated because three of the four items that compose the domain had no variability in scores between participants. This lack of variability is explained by the fact that because all participants were in a treatment programme within a prison, they all automatically received scores of 4 for Sexual Offence-Specific Treatment, External Control, and Supervised Living.

Further, there was only one numerically scored case for the Current Medication item so for all other cases, N/A scores for this item were replaced with zero when calculating totals. The ICC for the Professionally Provided Support domain for the Future context was .90, 95% CI [.77, 96], indicating excellent reliability. It must be considered however, that there was only one numerically scored case for the Future Medication item. For this reason, because Therapeutic Alliance was omitted for all cases, comparisons between Professionally Provided Support ICCs in the present study and those found in Willis et al. (2020) should be made with caution. Further, the Retrospective Scoring Guide utilised in the current study likely inflated ICCs

The percentage of cases in which raters were within 1 point of agreement was above 90% for all items for both Current and Future contexts with three exceptions. These were Leisure Activities for the Current context (no valid cases), Emotional Connection to Adults for the Current context (86%), and the Sexual Self-Regulation item in which agreement within 1 point was only found in two of the three valid cases. The average percent agreement across all items was 94.7% for Current context ratings and 94.4% for Future context ratings. For the medication item, both coders assigned the same participant a 4, and assigned every other participant N/A.

Predictive Validity

Over the follow-up period (*M* = 12.24 years), 12.9% (*n* = 27) of participants received convictions for a new sexual offence, 12.9% (*n* = 27) for a new violent offence, and 36.7% (*n* = 77) for a new general offence (non-sexual and non-violent). Of cases that reoffended, the average time between release and reoffending was 4.03 years (*SD* = 2.83; range = 36 days to 10.15 years) for sexual, 4.07 years (*SD* = 2.67; range = 33 days to 11.69 years) for violent, and 4.48 years (*SD* = 3.85; range = 7 days to 14.14 years) for general recidivism.

The primary aim of the current study was to assess the predictive validity of the SAPROF-SO. Predictive validity was assessed using ROC analyses. In the present study, SAPROF-SO AUCs can be interpreted as the probability that a randomly selected recidivist received a lower score than a randomly selected non-recidivist (Fawcett, 2006). Conversely, for the Static-99R, AUCs can be interpreted as the probability that a randomly selected recidivist received a higher score than a randomly selected non-recidivist. Results of ROC analyses assessing the predictive validity of SAPROF-SO scales (Current and Future) and the Static-99R for sexual, violent, and general recidivism are presented in Table 2.

It was found Total and Personal scores for both contexts were significantly predictive of sexual, violent, and general offending. These scales were most predictive of sexual recidivism, and of violent and general recidivism to a lesser extent. Total scores for both the Current and Future contexts produced large AUC values for sexual recidivism, AUC = .81, $p < .001$, and AUC = .78, $p < .001$, respectively. This was also true of Personal domain scores, AUC = .81, $p < .001$, and AUC = .80, $p < .001$. AUCs for SAPROF-SO Total scores were larger than the AUC for Static-99R scores in the prediction of sexual recidivism (AUC = .74, $p < .001$), although confidence intervals were overlapping. Compared to Total Current and Future scores, Static-99R scores were equally as predictive of violent recidivism and superior in the prediction of general offending.

In the prediction of sexual recidivism, Personal and Total scores for the Current context produced higher AUCs than Future context scores, but this difference was very small. There was almost no difference between Current and Future context ratings in the prediction of violent and general recidivism. Professionally Provided Support scales were not predictive of any type of offending.

Table 2

Predictive Validity of SAPROF-SO Scale Scores (Current and Future) and Static-99R Scores for Sexual, Violent, and General Recidivism

Scale	Sexual recidivism		Violent recidivism		General recidivism	
	AUC	AUC 95% CI	AUC	AUC 95% CI	AUC	AUC 95% CI
Personal – Current	.81***	[.72,.91]	.66**	[.54,.78]	.63**	[.55,.71]
Personal - Future	.80***	[.70,.90]	.65*	[.53,.78]	.63**	[.55,.71]
Professionally Provided Support – Current	.45	[.32,.57]	.51	[.40,.63]	.51	[.43,.59]
Professionally Provided Support – Future	.48	[.36,.60]	.45	[.35,.56]	.49	[.41,.57]
Total – Current	.81***	[.72,.90]	.66**	[.54,.79]	.63**	[.56,.71]
Total - Future	.78***	[.68,.87]	.65*	[.52,.77]	.63**	[.55,.70]
Static-99R	.74***	[.64,.85]	.66**	[.56,.77]	.68***	[.61,.76]

Note. SAPROF-SO = Structured Assessment of PROtective Factors – Sexual Offence version (Willis et al., 2017-2019). Static-99R (Helmus, Thornton et al., 2012). AUC = Area Under the Curve, CI = Confidence Interval. Missing scores were replaced with case mean domain scores in calculation of domain and total scores. Not applicable ratings were replaced with scores of 0.

*p <.05, **p<.01, ***p<.001

Incremental Validity

The current research aimed to establish whether and to what extent SAPROF-SO scales could provide incremental validity in predicting sexual recidivism after controlling for Static-99R scores. To achieve this aim, a series of hierarchical Cox regressions (proportional hazards model) were conducted. Cox regressions were selected as they allow the researcher to control for differences in time at large. Results of the Cox regression analyses are presented in Table 3. Static-99R scores were entered into the prediction model in Step 1 (Model 1) and SAPROF-SO Total and Personal scores for both Current and Future contexts were entered in Step 2, in four separate analyses (Models 2-5). Because the Professionally Provided Support domain was not predictive of reoffending, incremental validity analyses for this domain were not conducted.

As shown in Table 3, an initial analysis found Static-99R scores significantly predicted sexual recidivism. Hierarchical results showed that SAPROF-SO Total Current scores significantly predicted sexual recidivism after controlling for Static-99R scores and significantly improved the prediction model, $\Delta\chi^2(1) = 15.91$, $p < .001$. Further, it was found SAPROF-SO Total Future scores significantly predicted sexual recidivism after controlling for Static-99R scores and significantly improved the prediction model, $\Delta\chi^2(1) = 12.99$, $p < .001$. In both models, Static-99R scores retained significant unique predictive relationships to sexual recidivism.

Additionally, SAPROF-SO Personal Current scores significantly predicted sexual recidivism after controlling for Static-99R scores, and significantly improved the prediction model, $\Delta\chi^2(1) = 17.25$, $p < .001$. SAPROF-SO Personal Future scores significantly predicted sexual recidivism after controlling for Static-99R scores, and significantly improved the prediction model, $\Delta\chi^2(1) = 16.02$, $p < .001$. Static-99R scores retained significant unique predictive relationships to sexual recidivism in both models. In summary, results from the

current study supported the incremental validity of the SAPROF-SO Total and Personal scales over and above the Static-99R.

Table 3

Incremental Validity of SAPROF-SO Total and Personal Scores for both Contexts after Controlling for Static-99R Scores for Sexual Recidivism

Model	Regression coefficient			Hazard ratio		Model fit and improvement	
	B	*SE*	Wald	Exp(*B*)	95% CI	χ^2	$\Delta\chi^2(1)$
Model 1							
Step 1						25.42***	22.22***
Static-99R	0.34	.07	24.26***	1.41	[1.23,1.61]		
Model 2							
Step 2						37.96***	15.91***
Static-99R	0.19	.07	6.19*	1.20	[1.04,1.40]		
SAPROF-SO Total Current	-0.12	.03	14.01***	0.89	[0.83,0.94]		
Model 3							
Step 2						36.16***	12.97***
Static-99R	0.23	.07	10.38**	1.26	[1.09,1.44]		
SAPROF-SO Total Future	-0.09	.03	11.95**	0.92	[0.87,0.96]		
Model 4							
Step 2						39.14***	17.25***
Static-99R	0.18	.07	5.70*	1.20	[1.03,1.38]		
SAPROF-SO Personal Current	-0.13	.03	14.99***	0.89	[0.88,0.93]		
Model 5							
Step 2						37.95***	16.02***
Static-99R	0.20	.07	7.70**	1.22	[1.06,1.41]		
SAPROF-SO Personal Future	-0.11	.03	14.06***	0.90	[0.85,0.95]		

Note: SAPROF-SO = Structured Assessment of PROtective Factors – Sexual Offence version (Willis et al., 2017-2019). Static-99R (Helmus, Thornton et al., 2012). $\Delta\chi^2$ = change in chi square from previous step. CI = Confidence interval. Step 1 χ^2 *df* = 1. Step 2 χ^2 *df* = 2. Static-99R scores were entered into the prediction model in Step 1 (Model 1), and SAPROF-SO Total and Personal scores for both Current and Future contexts were entered in Step 2, in four separate analyses (Models 2-5).

*p <.05, **p<.01, ***p<.001

Discussion

To the author's knowledge, the SAPROF-SO is the first and only assessment tool that both exclusively assesses protective factors against sexual offending, and was developed for use with adults. The current study represents the first evaluation of the predictive validity of the SAPROF-SO. The present study utilised a retrospective design and primarily sought to examine the extent to which SAPROF-SO scores were predictive of sexual recidivism, using a sample of men convicted of sexual offences against chidren. The present study had a further aim to examine if the SAPROF-SO could provide incremental validity over and above scores from a well-validated measure of static risk (Static-99R). To aid in ruling out poor reliability as a threat to predictive accuracy, interrater reliability was analysed.

ICCs were excellent for SAPROF-SO Total, Personal, and Future context Professionally Provided Support scales. ICC analyses could not be run for Current context Professionally Provided Support scores due to very low variability in scores. This high level of consistency between raters suggests poor reliability was likely only a very weak source of predictive inaccuracy in the current study. It should be noted that the Retrospective Scoring Guide was developed to maximise interrater reliability by providing additional clarity around scoring each item. Therefore, the reported ICCs likely represent an overestimation of reliability relative to what might be expected in a professional setting or in a study that did not utilise a supplementary scoring guide. The use of the Retrospective Scoring Guide may explain why ICCs yielded from the current study were higher overall than those found in Willis et al. (2020).

Overall, results of the ROC analyses were supportive of the predictive validity of the SAPROF-SO. As hypothesised, SAPROF-SO Total scores for both Current and Future contexts were shown to be significantly predictive of sexual recidivism with AUCs falling in the large range (.81, .78). These AUC values were equivalent to those yielded by Beggs and

Grace (2010) in their validation of the VRS-SO using the same sample (albeit with eight additional participants in their study) and the same recidivism data. They reported AUCs of .79 and .80 for VRS-SO Total scores coded from both pre and posttreatment information, respectively. These results suggest that SAPROF-SO Total scores may be equally as predictive of sexual recidivism as VRS-SO Total scores among men who have been convicted of sexually offending against children. Given that results from Willis et al (2020) suggested the SAPROF-SO was capturing substantial variance independent of that captured by the VRS-SO, an important question for future research is whether that variance contributes incremental predictive validity. AUCs for SAPROF-SO Total scores were also higher than those for Static-99R scores in the prediction of sexual recidivism, although the confidence intervals were overlapping. This result suggests that the SAPROF-SO may be superior to the Static-99R in the prediction of sexual recidivism among men convicted of child sexual offences, but more research needs to be conducted. This result is perhaps surprising given the long follow-up period (*M*= 12.24 years) and the fact that the SAPROF-SO is largely comprised of dynamic factors that are susceptible to change (for better or worse) unlike the static factors within the Static-99R. Conversely, this result may be explained by the fact that SAPROF-SO scores were coded from posttreatment information which could theoretically be more informative in the prediction of recidivism than scores from static tools that would not change from pretreatment assessment, and therefore could not account for meaningful change achieved through therapy. This suggestion is supported by the similar finding from Beggs and Grace (2010) who found that the mostly dynamic VRS-SO was superior to the Static-99 in the prediction of sexual recidivism using the same sample. Together, these results highlight that tools with dynamic factors may be particularly preferable when assessing treated individuals.

In the prediction of sexual recidivism, AUCs for SAPROF-SO Total scores were larger than the AUC for SAPROF Total scores reported by de Vries Robbé, de Vogel, Koster et al. (2015) after their similar follow-up period of 15 years (AUC = .71). The difference between the AUCs yielded from these two studies may be explained by multiple factors. Their sample consisted of men with convictions for sexually violent offences and severe psychopathology, and predicting recidivism for men of this description may be more challenging than with men who have sexually offended against children. The discrepancy between AUCs between studies may be partially explained by the somewhat inferior interrater reliability of SAPROF Total scores (ICC = .85) found in their study. However, this difference in AUCs may also indicate the authors of the SAPROF-SO were successful in their effort to create a tool more specific to sexual offending than its predecessor. This final suggestion was further supported by the finding that SAPROF-SO Total and Personal scores were more predictive of sexual than violent or general reoffending, which was hypothesised. In addition, SAPROF-SO Total scores were significantly predictive of general and violent reoffending which likely reflects the overlap in protective factors relevant to these types of offending.

Results from the current study also supported the predictive validity of the Personal domain. Personal domain scores were significantly predictive of all three types of offending, and AUCs for sexual recidivism were large for both Current and Future context scores (.81, .80). Interestingly, the Professionally Provided Support scales were not significantly predictive of any type of offending. This result was likely influenced by a number of factors. This result may be explained by the suggestion that “the need for professionally provided support decreases with corresponding increases in protection from individual and social sources” (Willis et al., 2017-2019, p. 4). Theoretically, individuals with low scores on the Personal domain would tend to have high scores on the Professionally Provided Support

domain because their low level of protection necessitates high levels of intervention from professionals. This suggestion is supported by the finding from Willis et al. (2020) that Professionally Provided Support scores were significantly negatively correlated with DRAOR Protective scores (r = -.37, p < .05). Hypothetically, if a clinician assesses an individual as high in risk and/or low in protection they would recommend high levels of monitoring and restrictive sanctions, and the individual would score highly on Supervised Living and External Control. This is less of an issue with the Sexual Offence-Specific Treatment and Medication items because these items are coded less on the basis of "need" for intervention and more on the basis of other factors including availability, compliance and effectiveness (although "not required" and "not applicable" response options are available). Another explanation for this finding is that protection from professionally provided sources is often time-limited (e.g., time-limited prison sentences, parole conditions, mandated treatment etc.) and the current study utilised a follow-up period that may have continued long after the cessation of professionally provided support. In contrast, protective factors from within the Personal domain may have been more predictive of recidivism in the long term because they more strongly reflect protective personal characteristics that would theoretically be more enduring over time. Future research that uses a shorter follow-up period may yield larger AUC values for the Professionally Provided Support domain. It also likely that the poor predictive validity of the Professionally Provided Support domain was heavily influenced by the very low variability in scores between participants on the four items of which it was composed in this study. In a ROC analysis, AUC values will be low if there is limited variability in predictor variable data and there is adequate variability in extraneous predictor variables (e.g. Personal scores) that have a significant relationship with the outcome variable (e.g. recidivism). Because the participants were incarcerated and participating in the Kia Marama programme, all participants received the maximum score for Supervised Living,

External Control, and Sexual Offence-Specific Treatment for Current context ratings. There was also limited variability between participants for these items for Future context ratings, and a very high proportion of participants were scored N/A for the Medication item. Future research that utilises a sample with more variability in Professionally Provided Support scores may yield AUCs indicative of higher predictive accuracy. The decision to code N/A Medication scores as 0 (instead of 4 for example) may have negatively affected the predictive validity of Professionally Provided Support scores (and also Total scores by extension). This decision meant that those who did not require medication to lower their risk were scored lower than those who were not perfectly compliant with their required medication or were on medication that did not sufficiently reduce their risk (i.e., those that were scored 1, 2, or 3). Because only 5.7% of the current sample were numerically scored on this item, it is estimated this coding decision only marginally reduced the relevant AUCs yielded in this study. Regardless, in future studies, researchers may wish to revise how N/A scores are incorporated into score totals.

Interestingly, there was almost no difference in the AUCs for SAPROF-SO scores between Current and Future contexts across all three types of offending. It was predicted that Future context ratings would be more predictive than Current context ratings because many of the items that were coded separately are at least partially coded on the basis of protection present in the environment (e.g. Work, Leisure Activities, Housing Stability, Supervised Living, External Control, etc.) and Future context ratings may better reflect the environment in which the recidivism or desistance occurred (i.e. out of prison). However, Future context ratings (with the exception of one participant who remained in prison) were informed by postrelease plans pertaining to the upcoming 6-12 months, but for those that did reoffend, the average time between release and reoffending was over 4 years for each type of offending. Future context ratings therefore likely did not reflect the environment in which recidivism

occurred, and this result may have arisen because environmental sources of protection may be particularly sensitive to external influences and life events, and therefore may be too unstable to aid in the prediction of recidivism over the long-term. It should also be noted that Future context scores were based on discharge plans devised months before release which may or may not have eventuated, while Current context scores were based on more reliable information, such as observations from psychologists. It is possible the predictive validity of the Future context ratings was hampered by the reliability of the information used to inform them.

As hypothesised, the results of the current study also supported the incremental validity of the SAPROF-SO over and above the Static-99R. It was found in hierarchical analyses that SAPROF-SO Total and Personal scores for both Current and Future contexts significantly predicted sexual recidivism after controlling for Static-99R scores and significantly improved their respective prediction models in the four separate analyses. Further, in all analyses, Static-99R scores retained significant unique predictive relationships to sexual recidivism, which is perhaps unsurprising given the divergent validity analyses from Willis et al. (2020) and the fact that Static-99R has a focus on static risk factors while the SAPROF-SO has a focus on dynamic protective factors. The incremental validity findings support the suggestion that the SAPROF-SO is not merely measuring the same constructs as those found in static risk assessment tools but can provide additional predictive power. Further, the current results provide rationale for combining static risk and dynamic protective tools to increase the predictive accuracy of risk assessments (de Vries Robbé, de Vogel, Koster et al., 2015). However, as stated, further research must assess the incremental validity of the SAPROF-SO over and above measures of both static and dynamic risk, such as the VRS-SO.

Implications

Results from the current study were supportive of the predictive validity of the SAPROF-SO. Preliminary research has assessed construct validity (convergent and divergent), interrater reliability (Willis et al., 2020), and examined the factor structure of the tool (Thornton & Kelley, 2020). The current study represents a small step towards the tool meeting the requirements to be deemed actuarial and being used in forensic and correctional practice to inform risk predictions. It would be beneficial for future research to investigate the best method to combine scores, produce normative data, and investigate how scores should be divided into protection categories (e.g., Low Protection, Average Protection, High Protection, etc.; Willis et al., 2020; G. M. Willis, personal communication, December 22, 2020).

While it would be inappropriate at the present to use the SAPROF-SO as a tool to make predictions in regard to sexual recidivism, the tool may still be of use in professional settings in its current form. It is recommended that the SAPROF-SO be used alongside validated risk assessment tools to guide assessment which may identify strengths and factors that could be strengthened through intervention to lower risk (Willis et al., 2020). Further, the inclusion of proposed mechanisms in the coding manual may make the tool particularly useful in informing individual case formulations and treatment planning (Willis et al., 2020).

Assuming through further research the SAPROF-SO is found to make a useful contribution to actuarial risk assessment, the integration of the fully validated (and likely further refined) SAPROF-SO into professional practice will likely present many benefits. In theory, using the SAPROF-SO will lead to more balanced risk assessment and in particular could help to alleviate issues related to conducting deficit focussed risk evaluations (Cording & Beggs Christofferson, 2017; Miller, 2006; Rogers, 2000).

Additionally, introducing tools that measure protective factors such as the SAPROF-SO into clinical practice may lead to new directions and guidelines for treatment (de Ruiter & Nicholls, 2011). Through research that validates protective factors as treatment targets, new treatment modules and strategies may be created in order to target them. The SAPROF-SO may be especially useful to professionals that utilise strengths-based treatment frameworks, such as the GLM of which the SAPROF-SO is theoretically aligned (Willis et al., 2017-2019). The SAPROF-SO may enable clinicians to measure increases in strengths that may otherwise go undetected using tools that solely measure risk factors. Strength-based frameworks such as the GLM represent a promising way forward for sexual risk treatment, and could theoretically be superior to deficit focussed rehabilitation approaches such as the RNR model, in that strength building may more adequately motivate individuals with criminal histories to change their ways (Ward et al., 2007; Ward & Mann, 2004). The GLM has been criticised for lacking in empirical support (Mallion et al., 2020), however, the results of the current study are supportive of an underlying assumption of the GLM that strengths play a vital role in the prevention of harm to others.

Limitations and Future Research

One of the key limitations of the current study was the retrospective design. While the design afforded the opportunity to efficiently assess the predictive validity of the SAPROF-SO over a long follow-up period, the choice of design meant that coders were limited to information found in archived case files. Future research utilising a prospective design whereby coders could interview participants face to face, would allow researchers to gather all of the required information to score the SAPROF-SO, which would minimise omissions and provide predictive validity results that would more closely emulate the accuracy that could be expected if the SAPROF-SO was used in practice. Being restricted to case files presented a considerable challenge in the current study because protective factors, being a

newer concept to emerge in the field (Rogers, 2000), were not as commonly explicitly referred to in treatment documents as they are in more recent psychological practice. Using the Retrospective Scoring Guide was reasonably effective in minimising omissions and overcoming this limitation. However, the mean number of item omissions per participant was higher in the current study than what was found by Willis et al. (2020) in their high risk sample where scores were informed by far more recent file information (G. M. Willis, personal communication, February 18, 2021). In the present study, relevant information was particularly sparse in relation to Leisure Activities and Sexual Self-Regulation, leading to many item omissions. Further, information relating to the Therapeutic Alliance item was so minimal that this item was not included in the current study.

The use of the Retrospective Scoring Guide represents a threat to the external validity of the present study for two main reasons. The first is that it likely increased interrater reliability, as stated. The second is that the Retrospective Scoring Guide instructed coders to consider information that is not traditionally analysed when scoring the SAPROF-SO such as historical and psychometric information. Further, the coding rules for the Self-Control item were particularly different to those presented in the manual. It should be noted however that the Retrospective Scoring Guide was reviewed, critiqued, and amended by an author of the SAPROF-SO to ensure the content of the guide was theoretically aligned with the official manual. The amended coding rules for Attitudes Towards Rules and Regulations and Self-Control may have inflated AUC values in the current study. When there was insufficient information pertaining to recent functioning, the Retrospective Scoring Guide instructed raters to incorporate historical information including history of non-sexual offences into scoring decisions. Because historical non-sexual offences were incorporated into decision making, scores on these items may reflect the presence of a well-validated static risk factor (total number of prior offences; Hanson & Bussiere, 1998) as well as dynamic protective

factors, although the coder was instructed to give more weight to recent functioning. Future studies that are prospective or at least utilise more recent file information (and therefore do not necessitate the use of a supplementary scoring guide), may yield more externally valid results.

The homogenous nature of the sample presents a threat to the generalisability of the current findings. The current sample consisted only of men who volunteered for treatment and had convictions for sexually abusing children. It is possible that predictive validity results may have been different if the sample consisted of men who did not participate in voluntary treatment or had convictions for sexual offences against adults. Future research should measure the predictive validity of the SAPROF-SO using a more diverse sample of individuals. Another limitation of the current study was that reoffending was defined by reconvictions. Using reconvictions as a metric for sexually abusive behaviour is useful because this information is attainable from government databases and it would be very difficult if not impossible to attain reliable information about sexually abusive behaviours from the individuals themselves. For these reasons, reconviction data has been used in many other tool evaluation studies (Abbiati et al., 2017; Beggs & Grace, 2008; de Vries Robbé et al, 2013; de Vries Robbé, de Vogel, Koster et al., 2015; Turner et al., 2014). However, sceptics may interpret the predictive validity results as indicating that protective factors measured in the SAPROF-SO are negatively associated with getting apprehended for sexual offences, rather than sexually abusive behaviour.

The results of the current study showed that SAPROF-SO Total and Personal scores were significantly predictive of sexual reoffending, even after a long average follow-up time of 12.24 years. These results imply that the dynamic protective factors within the SAPROF-SO may be relatively stable over long time periods. Future research should assess the extent to which these protective factors are in fact malleable to change from treatment and how they

naturally change in the community. Researchers could conduct a study in which SAPROF-SO scores are coded at the beginning and end of treatment and change scores could be calculated to assess the extent to which the protective factors within the SAPROF-SO are malleable to change and therefore worth targetting through interventions. Additionally, researchers could conduct a study in which individuals are coded on the SAPROF-SO periodically after being released from prison to assess the natural pattern of change in the community. Such a study could shed light on the stability of protective factors in the commmunity and identify factors that strengthen or weaken protective factors such as life events (e.g. birth of a child, death of a family member etc.), which could be valuable in informing treatment and community-based risk management. In both suggested studies, researchers could then calculate if change scores are significantly predictive of reoffending after a follow-up period. Establishing that change scores are predictive of recidivism would strengthen the argument that the protective factors within the SAPROF-SO have a causal rather than purely correlational relationship with recidivism, which would further validate their use as treatment targets. Studies like those suggested have not yet been conducted using the SAPROF-SO, but a study with similar aims and methods has been conducted using the SAPROF. De Vries Robbé, de Vogel, Douglas et al. (2015) investigated the malleability of protective factors and the relationship between change scores and violent recidivism by retrospectively coding 108 men with violent and/or sexual convictions from information pertaining to both before and after treatment. They found SAPROF Total posttreatment scores were significantly higher than pretreatment scores, suggesting that protective factors can be strengthened through treatment. Further, they found that change in SAPROF Total, Internal and Motivational scores were significantly predictive of recidivism after both follow-up times of 1 and 11 years. Change in External domain scores were not predictive of recidivism after the 1 year follow-up, and this was likely because there are items within the

External domain that would theoretically be higher in those with low scores on the other domains (akin to the issue with the Professonally Provided Support domain in the current study). Taken together, the results of their study suggest that the protective factors within the SAPROF represent valuable treatment targets as they are malleable to treatment and may have a causal relationship with desistance from offending, yet these factors appear stable enough in the community to be useful in the prediction of recidivism in the long term.

It has been suggested that incorporating protective factors into assessment and treatment could increase treatment engagement and motivation, and improve the therapeutic alliance, which in turn could lead to reduced recidivism rates (de Vries Robbé, de Vogel, Koster et al., 2015, de Vries Robbé & Willis, 2017). Future research should endeavour to empirically validate this claim. For example, in a prospective study, researchers could measure changes in engagement, motivation and the quality of the therapeutic alliance throughout treatment using psychometrics such as the University of Rhode Island Change Assessment (McConnaughy et al., 1983), and the Working Alliance Inventory (Hovarth & Greenberg, 1989). Researchers could compare scores from a sample that received strength focused assessment and treatment to scores from a sample (matched on key variables) that were assessed and treated via a deficit focussed approach. Further, researchers could analyse which approach was associated with greater reductions in recidivism.

Finally, future research should investigate the predictive value of individual items within the SAPROF-SO. Researchers could investigate combining highly predictive items from the SAPROF-SO with highly predictive items from well-validated and actuarial static and dynamic risk assessment tools such as the Static-99R and the VRS-SO to create a “super tool” that could theoretically explain maximum variance in sexual recidivism outcomes. Because it can be a lengthy exercise to code individuals on multiple tools, a further benefit of

a "super tool" would be that it could potentially lead to more time-efficient and therefore cost-effective assessments of risk.

Conclusion

The results of the current study support the growing body of empirical evidence that suggests measuring protective factors can be valuable in predicting recidivism (de Vries Robbé, de Vogel, Koster et al., 2015; de Vries Robbé et al., 2011; de Vries Robbé, de Vogel, Douglas et al., 2015; Abbiati et al., 2017; Yoon et al., 2018). The current study is the first to both investigate and find evidence for the predictive validity of the SAPROF-SO. In summary, it was found that SAPROF-SO Total and Personal scores were significantly predictive of sexual, violent, and general offending after an average follow-up period of 12.24 years. Further, in the prediction of sexual recidivism, the obtained AUC values for Total and Personal scores were equivalent to or higher than values obtained from research evaluating well-validated and widely used assessment tools that rely on risk factors (Beggs & Grace, 2010; Brankley et al., 2019; Helmus, Hanson et al., 2012; Jackson, 2016; Kelley et al., 2020, Olver et al., 2007; Olver et al., 2014). The current results also support the incremental validity of the Total and Personal scales over and above a well-validated and widely used measure of static risk (Static-99R), and provide impetus to utilising multiple tools to capture both static risk and dynamic protective factors. It should be noted that the retrospective design necessitated the creation and use of a supplementary scoring guide to overcome the scarcity of relevant information in case files to score each item. While the Retrospective Scoring Guide was critiqued, amended and approved by an author of the SAPROF-SO, future studies that utilise more recent file information or use a prospective design may yield predictive validity results that more closely emulate what could be expected if the tool was to be used in professional practice.

Throughout the history of risk assessment it has been common practice to focus only on the negative aspects of those assessed, likely leading to unbalanced evaluations. The results of the current study provide rationale for attending to the positive.

References

Abbiati, M., Azzola, A., Palix, J., Gasser, J., & Moulin, V. (2017). Validity and predictive accuracy of the Structured Assessment of Protective Factors for Violence Risk in criminal forensic evaluations: A Swiss cross-calidation retrospective study. *Criminal Justice and Behavior*, *44*(4), 493-510. https://doi.org/10.1177/0093854816677565

Andrews, D. (1995) The Psychology of criminal conduct and effective treatment. In McGuire, J. (Ed.), *What works: Reducing re-offending: Guidelines from research and practice* (pp. 35-62). John Wiley & Sons Ltd.

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, *52*(1), 7-27. https://doi.org/10.1177/0011128705281756

Bakker, L. W., Riley, D., & O'Malley, J. (1999). *ROC, Risk of reconviction: Statistical models predicting four types of re-offending*. Department of Corrections. https://doc-web.squiz.cloud/__data/assets/pdf_file/0020/10667/roc.pdf

Beggs, S. M. (2008). Treatment outcome, risk assessment, and recidivism among sexual offenders against children [Doctoral dissertation, University of Canterbury]. UC Research Depository. https://ir.canterbury.ac.nz/bitstream/handle/10092/1824/thesis_fulltextpdf.pdf?sequence=1&isAllowed=y

Beggs, S. M., & Grace, R. C. (2008). Psychopathy, intelligence, and recidivism in child molesters: Evidence of an interaction effect. *Criminal Justice and Behavior*, *35*(6), 683-695. https://doi.org/10.1037/a0022900

Beggs, S. M., & Grace, R. C. (2010). Assessment of dynamic risk factors: an independent validation study of the Violence Risk Scale: Sexual Offender version. *Sexual Abuse, 22*, 234-251. https://doi.org/10.1177/1079063210369014

Beggs, S. M., & Grace, R. C. (2011). Treatment gain for sexual offenders against children predicts reduced recidivism: a comparative validity study. *Journal of Consulting and Clinical Psychology*, *79*, 182-192. https://doi.org/10.1037/a0022900

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20: Professional guidelines for assessing risk of sexual violence.* British Columbia Institute against Family Violence.

Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (p. 18-32). Sage Publications, Inc.

Bonta, J., & Andrews, D. A. (2017). *The psychology of criminal conduct* (6th ed.). Taylor & Francis.

Bonta, J., Wallace-Capretta, S., & Rooney, J. (2000). A quasi-experimental evaluation of an intensive rehabilitation supervision program. *Criminal Justice and Behavior*, *27*(3), 312-329. https://doi.org/10.1177/0093854800027003003

Brankley, A. E., Babchishin, K. M., & Hanson, R. K. (2019). STABLE-2007 demonstrates predictive and incremental validity in assessing risk-relevant propensities for sexual offending: A meta-analysis. *Sexual Abuse*, *33*(1), 34-63. https://doi.org/10.1177/1079063219871572

Brown, J. D. (1996). *Testing in Language Programs.* Prentice-Hall, Inc.

Caldwell, R. A., Bogat, G. A., & Davidson, W. S. (1988). The assessment of child abuse potential and the prevention of child abuse and neglect: A policy analysis. *American Journal of Community Psychology*, *16*(5), 609-624. https://doi.org/10.1007/BF00930017

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105. https://doi.org/10.1037/h0046016

Canales, D. D., Olver, M. E., & Wong, S. C. P. (2009). Construct Validity of the Violence Risk Scale—Sexual Offender Version for Measuring Sexual Deviance. *Sexual Abuse*, *21*(4), 474-492. https://doi.org/10.1177/1079063209344990

Cording, J. R., & Beggs Christofferson, S. M. (2017). Theoretical and practical issues for the measurement of protective factors. *Aggression and Violent Behavior*, *32*, 45-54. https://doi.org/10.1016/j.avb.2016.12.007

Department of Corrections (2009). *What works now? A review and update of research evidence relevant to offender rehabilitation practices within the Department of Corrections*. https://www.corrections.govt.nz/__data/assets/pdf_file/0004/12775/What_Works_Now_Final_December_2009.pdf

Department of Corrections (2011). *Reconviction rates of sex offenders. Five year follow up study: Sex offenders against children vs offenders against adults.* http://www.corrections.govt.nz/__data/assets/pdf_file/0008/672317/Reconviction_Rates_of_Sex_Offenders_-_five_year_follow_up.pdf

de Ruiter, C., & Nicholls, T. L. (2011). Protective factors in forensic mental health: A new frontier. *International Journal of Forensic Mental Health*, *10*(3), 160-170. https://doi.org/10.1080/14999013.2011.600602

de Vogel, V., de Ruiter, C., Bouman, Y., & de Vries Robbé, M. (2012). *SAPROF: Guidelines for the assessment of protective factors for violence risk* (2nd ed.). Van der Hoeven Stichting.

de Vogel, V., de Ruiter, C., van Beek, D., Mead, G. (2004). Predictive accuracy of the SVR-20 and Static-99 in a Dutch sample of treated sex offenders. *Law and Human Behavior*, *28*, 235-251. https://doi.org/10.1023/B:LAHU.0000029137.41974.eb

de Vries Robbé, M., de Vogel, V., & de Spa, E. (2011). Protective factors for violence risk in forensic psychiatric patients: A retrospective validation study of the SAPROF. *International Journal of Forensic Mental Health*, *10*(3), 178-186. https://doi.org/10.1080/14999013.2011.600232

de Vries Robbé, M., de Vogel, V., Douglas, K. S., & Nijman, H. L. (2015). Changes in dynamic risk and protective factors for violence during inpatient forensic psychiatric treatment: Predicting reductions in postdischarge community recidivism. *Law and Human Behavior*, *39*(1), 53-61. https://doi.org/10.1037/lhb0000089

de Vries Robbé, M., de Vogel, V., Koster, K., & Bogaerts, S. (2015). Assessing protective factors for sexually violent offending with the SAPROF. *Sexual Abuse*, *27*(1), 51-70. https://doi.org/10.1177/1079063214550168

de Vries Robbé, M., Mann, R. E., Maruna, S., & Thornton, D. (2015). An exploration of protective factors supporting desistance from sexual offending. *Sexual Abuse*, *27*(1), 16-33. https://doi.org/10.1177/1079063214547582

de Vries Robbé, M., & Willis, G. M. (2017). Assessment of protective factors in clinical practice. *Aggression and Violent Behavior*, *32*, 55-63. https://doi.org/10.1016/j.avb.2016.12.006

Doren, D. M. (2006). Recidivism risk assessments: Making sense of controversies. In W. L. Marshall, Y. M. Fernandez, L. E. Marshall, & G. A. Serran (Eds.), *Sexual offender treatment: Controversial issues* (pp. 3-15). Wiley.

Douglas, K. S., & Kropp, P. R. (2002). A prevention-based paradigm for violence risk assessment: Clinical and research applications. *Criminal Justice and Behavior*, *29*(5), 617-658. https://doi.org/10.1177/009385402236735

Douglas, K. S., & Reeves, K. A. (2010). Historical-Clinical-Risk Management-20 (HCR-20) Violence Risk Assessment Scheme: Rationale, application, and empirical overview. In R. K. Otto & K. S. Douglas (Eds.), *International perspectives on forensic mental health. Handbook of violence risk assessment* (pp. 147-185). Taylor & Francis.

Eher, R., Olver, M. E., Heurix, I., Schilling, F., & Rettenberger, M. (2015). Predicting reoffense in pedophilic child molesters by clinical diagnoses and risk assessment. *Law and Human Behavior*, *39*(6), 571-580. https://doi.org/10.1037/lhb0000144

Farrington, D. P., Ttofi, M. M., & Piquero, A. R. (2016). Risk, promotive, and protective factors in youth offending: Results from the Cambridge study in delinquent development. *Journal of Criminal Justice*, *45*, 63-70. https://doi.org/10.1016/j.jcrimjus.2016.02.014

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.

Hanby, L. J. (2013). A longitudinal study of dynamic risk, protective factors, and criminal recidivism: Change over time and the impact of assessment timing [Doctoral dissertation, Carleton University]. Carleton University Research Virtual Environment. https://curve.carleton.ca/b6aa7d5d-f44b-48d6-9ea9-5d0f92b8f7a7

Hanson, R. K. (2009). The psychological assessment of risk for crime and violence. *Canadian Psychology, 50*(3), 172-182. https://doi.org/10.1037/a0015726

Hanson, R. K., Babchishin, K. M., Helmus, L. M., Thornton, D., & Phenix, A. (2017). Communicating the results of criterion referenced prediction measures: Risk categories for the Static-99R and Static-2002R sexual offender risk assessment tools. *Psychological Assessment, 29*(5), 582-597. https://doi.org/10.1037/pas0000371

Hanson, R. K., & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66*(2), 348-362. https://doi.org/10.1037/0022-006X.66.2.348

Hanson, R. K., Harris, A. J., Scott, T. L., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The Dynamic Supervision Project*. Public Safety Canada. https://ccoso.org/sites/default/files/import/risk-assessment.pdf

Hanson, R. K., Lunetta, A., Phenix, A., Neeley, J., & Epperson, D. (2014). The field validity of Static-99/R sex offender risk assessment tool in California. *Journal of Threat Assessment and Management, 1*(2), 102-117. https://doi.org/10.1037/tam0000014

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21*(1), 1-21. https://doi.org/10.1037/a0014421

Harris, A., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static 99: Coding rules revised 2003*. Solicitor General Canada. http://www.defenseforsvp.com/Resources/Hanson_Static-99/Static99codingRules_e.pdf

Haynes, S. N., & Lench, H. C. (2003). Incremental Validity of New Clinical Assessment Measures. *Psychological Assessment, 15*(4), 456-466. https://doi.org/10.1037/1040-3590.15.4.456

Helmus, L., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. (2012). Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: A meta-analysis. *Criminal Justice and Behavior*, *39*(9), 1148-1171. https://doi.org/10.1177/0093854812443648

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse*, *24*(1), 64-101. https://doi.org/10.1177/1079063211409951

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*(4), 446-455. https://doi.org/10.1037/1040-3590.15.4.446

Jackson, K. J. (2016). Validation of the Risk for Sexual Violence Protocol in adult sexual offenders [Doctoral dissertation, Simon Fraser University]. Summit: Simon Fraser University Research Depository. http://summit.sfu.ca/item/16622

Kelley, S. M., Ambroziak, G., Thornton, D., & Barahal, R. M. (2020). How do professionals assess sexual recidivism risk? An updated survey of practices. *Sexual Abuse*, *32*(1), 3-29. https://doi.org/10.1177/1079063218800474

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Lowenkamp, C. T., & Latessa, E. J. (2004). Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in Community Corrections, 2004*, 3-8. https://www.researchgate.net/publication/309457017

Mallion, J. S., Wood, J. L., & Mallion, A. (2020). Systematic review of 'Good Lives' assumptions and interventions. *Aggression and Violent Behavior*, *55*, Article 101510. https://doi.org/10.1016/j.avb.2020.101510

Mann, R. E., Hanson, R. K., & Thornton, D. (2010). Assessing risk for sexual recidivism: Some proposals on the nature of psychologically meaningful risk factors. *Sexual Abuse*, *22*(2), 191-217. https://doi.org/10.1177/1079063210366039

McConnaughy, E. A., Prochaska, J. O., & Velicer, W. F. (1983). Stages of change in psychotherapy: Measurement and sample profiles. *Psychotherapy: Theory, Research & Practice, 20*(3), 368-375. https://doi.org/10.1037/h0090198

McGrath, R. J., Lasher, M. P., & Cumming, G. F. (2012). The Sex Offender Treatment Intervention and Progress Scale (SOTIPS): Psychometric properties and incremental predictive validity with Static-99R. *Sexual Abuse: A Journal of Research and Treatment, 24*(5), 431-458. https://doi.org/10.1177/1079063211432475

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282. https://hrcak.srce.hr/89395

Messick, S. (1989). *Validity.* In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (p. 13-103). Macmillan Publishing Co, Inc.; American Council on Education.

Miller, H. A. (2006). A dynamic assessment of offender risk, needs, and strengths in a sample of pre- release general offenders. *Behavioral Sciences & the Law*, *24*(6), 767-782. https://doi.org/10.1002/bsl.728

Ministry of Justice. (2016a). *Research and data: sexual violence.* https://www.justice.govt.nz/justice-sector-policy/research-data/nzcass/survey-results/results-by-subject/sexual-violence/

Ministry of Justice. (2016b). *Sex offender treatment for adults: evidence brief*. https://www.justice.govt.nz/assets/Documents/Publications/Sex-Offender-Treatment-for-Adults.pdf

Murphy, K. R., & Davidshofer, C. O. (2004). *Psychological testing: Principles and applications* (6th ed.). Pearson.

Murrie, D. C., Boccaccini, M. T., Turner, D. B., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis) agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law*, *15*(1), 19-53. https://doi.org/10.1037/a0014897

Nadesu, A. (2011). *Reconviction rates of sex offenders: Five year follow-up study: Sex offenders against children vs offenders against adults*. Department of Corrections. https://www.corrections.govt.nz/__data/assets/pdf_file/0016/10708/Reconviction_Rates_of_Sex_Offenders_-_five_year_follow_up.pdf

Olver, M. E., Nicholaichuk, T. P., Kingston, D. A., & Wong, S. C. (2014). A multisite examination of sexual violence risk and therapeutic change. *Journal of Consulting and Clinical Psychology*, *82*(2), 312-324. https://doi.org/10.1037/a0035340

Olver, M. E., Wong, S. C., Nicholaichuk, T., & Gordon, A. (2007). The validity and reliability of the Violence Risk Scale-Sexual Offender version: Assessing sex offender risk and evaluating therapeutic change. *Psychological Assessment*, *19*(3), 318-329. https://doi.org/10.1037/1040-3590.19.3.318

Papalia, N., Spivak, B., Daffern, M., & Ogloff, J. R. (2020). Are psychological treatments for adults with histories of violent offending associated with change in dynamic risk factors? A meta-analysis of intermediate treatment outcomes. *Criminal Justice and Behavior*, *47*(12), 1585-1608. https://doi.org/10.1177/0093854820956377

Reeves, S. G., Ogloff, J. R., & Simmons, M. (2018). The predictive validity of the Static-99, Static-99R, and Static-2002/R: Which one to use? *Sexual Abuse*, *30*(8), 887-907. https://doi.org/10.1177/1079063217712216

Rettenberger, M., Boer, D. P., & Eher, R. (2011). The Predictive Accuracy of Risk Factors in the Sexual Violence Risk-20 (SVR-20). *Criminal Justice and Behavior*, *38*(10), 1009-1027. https://doi.org/10.1177/0093854811416908

Rettenberger, M., Hucker, S. J., Boer, D. P., & Eher, R. (2009). The Reliability and Validity of the Sexual Violence Risk-20 (SVR-20): An International Review. *Sexual Offender Treatment*, *4*(2), 1-14. https://hdl.handle.net/10289/6305

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, *63*(5), 737-748. https://doi.org/10.1037/0022-006X.63.5.737

Rogers, R. (2000). The uncritical acceptance of risk assessment in forensic practice. *Law and Human Behavior*, *24*(5), 595-605. https://doi.org/10.1023/A:1005575113507

Serin, R. C. (2007). *The Dynamic Risk Assessment Scale for Offender Re-Entry (DRAOR)*. Carleton University.

Siedlecki, S. L., & Albert, N. M. (2017). Understanding interrater reliability and validity of risk assessment tools used to predict adverse clinical events. *Clinical Nurse Specialist*, *31*(1), 23-29. https://doi.org/10.1097/NUR.0000000000000260

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical psychology review*, *31*(3), 499-513. https://doi.org/10.1016/j.cpr.2010.11.009

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science*, *20*(1), 38-42. https://doi.org/10.1177/0963721410397271

Skelton, A., & Vess, J. (2008). Risk of sexual recidivism as a function of age and actuarial risk. *Journal of Sexual Aggression*, *14*(3), 199-209. https://doi.org/10.1080/13552600802267098

Sowden, J. N. (2013). Examining the relationship of risk, treatment readiness, and therapeutic change to recidivism in a sample of treated sex offenders [Doctoral dissertation, University of Saskatchewan]. Harvest: University of Saskatchewan's Research Archive. https://harvest.usask.ca/handle/10388/ETD-2013-11-1305

Sowden, J. N., & Olver, M. E. (2017). Use of the Violence Risk Scale—Sexual Offender Version and the Stable 2007 to assess dynamic sexual violence risk in a sample of

treated sexual offenders. *Psychological Assessment*, *29*(3), 293-303. https://doi.org/10.1037/pas0000345

Stadtland, C., Hollweg, M., Kleindienst, N., Dietl, J., Reich, U., & Nedopil, N. (2005). Risk assessment and prediction of violent and sexual recidivism in sex offenders: Long-term predictive validity of four risk assessment instruments. *Journal of Forensic Psychiatry & Psychology*, *16*(1), 92-108. https://doi.org/10.1080/1478994042000270247

Thornton, D., & Kelley, S. M. (2020, October 21-23). *Developing protective factors responsive to different criminogenic need profiles* [Conference session]. Association for the Treatment of Sexual Abusers (ATSA) Annual Research and Treatment Conference. https://www.atsa.com/Public/Conference/2020/ATSA_2020_Conference_Program.pdf

Thornton, D., Kelley, S. M., Nelligan, K. E. (2017). Protective factors and mental illness in men with a history of sexual offending. *Aggression and Violent Behavior*, *32*, 29-36. https://doi.org/10.1016/j.avb.2016.12.003

Ward, T., & Mann, R. (2004). Good lives and the rehabilitation of offenders: A positive approach to sex offender treatment. In Linley, P. A., Joseph, S. (Eds.), *Positive Psychology in Practice* (pp. 598-616). Wiley.

Ward, T., Melser, J., & Yates, P. M. (2007). Reconstructing the Risk-Need-Responsivity model: A theoretical elaboration and evaluation. *Aggression and Violent Behavior*, *12*(2), 208-228. https://doi.org/10.1016/j.avb.2006.07.001

Ward, T., & Stewart, C. A. (2003). The treatment of sex offenders: Risk management and good lives. *Professional Psychology: Research and Practice*, *34*(4), 353-360. https://doi.org/10.1037/0735-7028.34.4.353

Willis, G. M., & Grace, R. C. (2008). The quality of community reintegration planning for child molesters: Effects on sexual recidivism. *Sexual Abuse*, *20*(2), 218-240. https://doi.org/10.1177/1079063208318005

Willis, G. M., & Grace, R. C. (2009). Assessment of community reintegration planning for sex offenders: Poor planning predicts recidivism. *Criminal Justice and Behavior*, *36*(5), 494-512. https://doi.org/10.1177/0093854809332874

Willis, G. M., Kelley, S. M., & Thornton, D. (2020). Are protective factors valid constructs? Interrater reliability and construct validity of proposed protective factors against sexual reoffending. *Criminal Justice and Behavior*, *47*(11), 1448-1467. https://doi.org/10.1177/0093854820941039

Willis, G. M., Thornton, D. T., Kelley, S. M., & de Vries Robbé, M. (2017-2019). *The Structured Assessment of Protective Factors for Violence Risk - Sexual Offending (SAPROF-SO). Pilot manual.*

Wong, S., Olver, M. E., Nicholaichuk, T. P., & Gordon, A. (2003-2017). *The Violence Risk Scale—Sexual Offender version (VRS-SO).* Regional Psychiatric Centre and University of Saskatchewan.

Worling, J. R. (2013). *Desistence for adolescents who sexually harm*. Autor. http://www.drjamesworling.com/uploads/8/7/7/6/8776493/dash-13.pdf

Zeng, G., Chu, C. M., & Lee, Y. (2015). Assessing protective factors of youth who sexually offended in Singapore: Preliminary evidence on the utility of the DASH-13 and the SAPROF. *Sexual Abuse, 27*(1), 91-108. https://doi.org/10.1177/1079063214561684

Appendix A

Table A1

Descriptions of SAPROF-SO Domains and Items

SAPROF-SO domains and items	Description
Personal	These items relate to: internal protective capacities; having a prosocial and/or adaptive identity; the extent to which an individual is connected to prosocial others and/or activities; or having housing and financial stability.
1. Intact cognitive functioning	Considers intelligence and general cognitive abilities at baseline and the past six months.
2. Secure attachment in childhood	Considers the presence of a close, warm, loving relationship with at least one prosocial adult in the first 18 years of life.
3. Adaptive schemas	Considers global representation of the self and others that have been adaptive for the past year and will likely continue to be adaptive.
4. Empathy	Considers the person's ability to be able to take the perspective of others and engage in helpful responding.
5. Coping	Considers whether the person is managing general life stressors in effective ways.
6. Self-control	Considers whether the person is managing impulses and delaying immediate gratification.
7. Sexual self-regulation	Considers the regulation of sexual impulses and evidence of a normative sex drive.
8. Prosocial sexual interests	Considers whether the person has an interest in, and arousal to, consenting adult sex.
9. Prosocial sexual identity	Considers whether the person is accepting of their prosocial adult sexual orientation.
10. Goal-directed living	Considers whether the individual has prosocial, meaningful goals for living that drive prosocial behavior.
11. Motivation for managing risk	Considers the motivation to manage risk factors associated with sexual offending.
12. Attitudes towards rules and regulations	Considers one's acceptance of the importance of rules/regulations and willingness to comply with them.
13. Work	Considers stable and suitable paid or voluntary work that is intrinsically motivating.
14. Leisure activities	Considers the person's engagement in structured, enjoyable activities with prosocial others.

SAPROF-SO domains and items	Descriptions
15. Social network	Considers whether the individual has a prosocial and supportive group of people who are not paid to be with them.
16. Emotional connection to adults	Considers whether the individual has established emotionally intimate bonds with other adults.
17. Intimate relationship	Considers whether the individual has a stable romantic relationship of good quality, which includes physical intimacy.
18. Housing stability	Considers the individual's access to stable accommodation.
19. Financial management	Considers whether the individual has a steady income and financial management skills.
Professionally Provided Support	This domain consists of five items that pertain to additional support provided by external sources
20. Sexual offence-specific treatment	Considers the availability of appropriate treatment services.
21. Medication	Considers the individual's compliance and motivation to take prescribed medication as well as whether the medication is effective in reducing symptoms directly or indirectly associated with sexual recidivism risk (e.g., hypersexuality).
22. Therapeutic alliance	Considers the individual's perception of a warm, positive therapeutic or supervisory alliance.
23. Supervised living	Considers the extent to which an individual's living situation is formally or informally supervised.
24. External control	Considers the intensity of court-ordered or mandatory supervision and/or treatment.

Note. SAPROF-SO = Structured Assessment of PROtective Factors – Sexual Offence version. Adapted from the supplemental materials of "Are protective factors valid constructs? Interrater reliability and construct validity of proposed protective factors against sexual reoffending," by G. M. Willis, S. M. Kelley, and D. Thornton, 2020, *Criminal Justice and Behavior*, *47*(11), pp. 1-2 (https://doi.org/10.25384/SAGE.12667146.v1). Copyright 2020 by the International Association for Correctional and Forensic Psychology.

Appendix B

The Retrospective Scoring Guide

Guide to scoring the SAPROF-SO from Kia Marama case files

General guidelines

This guide is designed to give direction on how to score the SAPROF-SO using the unique information in Kia Marama case files from the 1990s. It is to be used alongside the official SAPROF-SO scoring manual. The first aim is to direct the coder where to find relevant information within the files. An issue with performing retrospective coding from 1990s case files is that protective factors were often not explicitly mentioned. This guide aims to minimise item omissions by allowing coders to integrate information that is not typically considered such as historical information and scores from relevant psychometrics. The coding guide also gives more explicit coding rules and aims to maximise interrater reliability.

Setting up the coding form

Before you start, it is useful to write the following at the top of the page:

Date of release:	(Usually the month of the most recent report)
Date of sentencing:	(Found in Notes of Sentencing or Presentence Report)
Dates of CSA offences:	(Found in Summary of Facts or Notes of Sentencing)
Time between CSA and sentencing:	(To nearest month)
DOB:	(Found in Demographic Questionnaire, Report etc.)
Age when offences occurred:	
Adolescent limited: Y/N	

As you read through documents in the file it is useful to attain this information as it will speed up the process of coding items that have set timeframes in their criteria (e.g. Prosocial Sexual Interests etc.).

Finding and removing documents

Next, go through each file and check to see if the documents in the following list are in each file and remove them from the file. If it is a bound and organized file with clips, make note of where each document is kept so you can later return the file to its original state. The format for how each case file is organised seems to change from year to year so I cannot reliably tell you where each document will be – you will have to sift through. What is included in each case file changes from year to year, so I also cannot tell you whether each document will be in the file. Again, you will have to sift through.

1. Find the longest “Psychological Report” in the file. This will usually be the most recent report. Every case file will have one of these and the longest ones are usually to “Community Corrections”. If there is another report for “Psychological Services” also remove this as there may be additional information in the back, usually pertaining to further psychological work to be done.

2. Find the “Demographic Questionnaire”, this is typically found near the front of the file and there is one of these in every file.
3. Next, find the “Test Score Summary Sheet”. It is usually found at the start of the psychometric section which usually starts about halfway through. In rare cases this will be missing.
4. Find the documents: “Rating Scale – Pre Treatment” and “Rating Scale – Post Treatment”. These are usually found together, sometimes at the front and sometimes near the “Test Score Summary Sheet”. These are scoring sheets are only found in roughly one third of files.
5. WAIS-R Scoring form, this will be among the psychometrics. This is useful for the first item.
6. Find the Hare PCL-R scoring form among the psychometrics. This is filled out by the therapist post treatment so is full of relevant and reliable information. It is important to remember this form aims to identify risk factors rather than protective factors.
7. Find the “Summary of Facts” or if this is not present then the “Sentencing Notes” among the legal paperwork. This is useful because you will need to know *when* the offences took place relative to sentencing. Knowing the age of the victims is also relevant but you do not need to read the details of the crimes themselves.
8. Find the “Presentence Report”. This is usually found among the legal paperwork. This document is vital because it contains a section on the financial situation of the person which is used to code Financial Management. It also contains information on their response to previous sentences which is useful for coding Attitudes Towards Rules and Regulations and to some extent Self Control. It should be read in its entirety as it is often useful for coding Internal Capacity Items.
9. Some men will have a document called “My Personal Statement” among their therapy work which can be useful in showing strategies in avoiding future offending and other relevant information. It is important to remember these strategies are mainly fed to the prisoners by their therapist so whether you believe they will adhere to them upon release is up to your discretion.
10. Find the “TPQ” among the psychometrics. There are two questions with relevant answers to the Attitudes Towards Rules and Regulations. This is often not present in files after 1995 as they began to use the MCMI.
11. “Social History” – This is a handwritten document that is typically found among their therapy work. This is useful for coding Secure Attachment (if you can read it).
12. “Sexual History” – This is useful for coding Prosocial Sexual Interests and Sexual Self-Regulation.
13. “Relapse Prevention Interview” –This is useful for rating element 2 of Sexual Self-Regulation. This document is routinely missing.

This should be all you need to code a participant but if you may find other relevant documents as you search through the file. The MCMI can be useful to code Internal Capacity Items if you are familiar with the tool. There also may be psychological reports from during or before their time in prison and there may be reports from probation officers from previous sentences.

Recommended Procedure

Once you are familiar with the official scoring manual and what information is relevant to each item, it is recommended you read these documents in the following order and write the information into the relevant boxes on the coding form as you go. It is recommended you only write scores once you have read the full report at the end.

1. Summary of Facts/Sentencing Notes – Identify age and gender of victim(s), date(s) of CSA offences and sentencing, and whether offending was adolescent limited.
2. Presentence Report – Identify financial situation, response to previous charges, look for other relevant info.
3. Demographic Questionnaire – There is a lot of useful information, particularly about their childhood, preferred victim age, work stability, alcohol and drug use etc.
4. Test Score Summary Sheet – In the relevant boxes in the form record whether they have clinically significant pre-treatment **AND** post-treatment scores (or not) on the BDI, SSEI, FIS, UCLS, HTW (Adaptive Schema), STAI – Trait, STAXI – Trait, STAXI-Suppression (Coping), STAXI-Expression, STAXI-Suppression, STAXI-Control (Self-control).
5. If you did not find their WAIS-R IQ score in the Demographic Questionnaire or Test Score Summary Sheet check for this on the WAIS-R form.
6. Hare PCL-R – Record their scores for: Grandiosity (Adaptive Schema), Lack of Empathy, Cunning and Manipulative (Empathy), Poor Behavioural Controls (Coping & Self-Control), Impulsivity (Self-Control), Lack of Realistic Long-term Goals (Goal-Directed Living), Parasitic Lifestyle (Financial Management) Revocation of Conditional Release and Criminal Versatility (Attitudes Towards Rules and Regulations & Self Control).
7. TPQ – Record their answers to items 13 and 24 (Attitudes Towards Rules and Regulations).
8. Rating Scale – Pre and Post Treatment – Record their scores for Stress Management and Problem Solving (Coping), Sexual Preoccupation (Sexual Self-Regulation) and Motivation for Change (Motivation for Managing Risk).
9. My Personal Statement – Information in this document (if present) is useful for coding element 1 and 2 of Sexual Self-Regulation and Coping.
10. Social History – This is useful for coding Secure Attachment.
11. Sexual History – This is useful for coding Prosocial Sexual Interests.
12. Relapse Prevention Interview - This is useful for rating element 2 of Sexual Self-Regulation.
13. Report to Community Corrections and/or Psychological Services – As you read this, you can start scoring each item using criteria from the official scoring manual and the guidelines below.

Scoring guidelines

For all **Internal Capacity** and **Prosocial Identity** items, copy the "current context" score you decide on into the "future context" box. For the remainder of the items, score them separately.

Due to the lack of information in these files about relationships between participants and their therapists, you do not code Item 22 – Therapeutic Alliance.

Reminder: If you are deciding between two anchor points (0,2,4) and cannot make a decision, score the number between those anchor points. When deciding between an anchor point and an in between point (1,3), score the anchor point.

The following section aims to give clarity on what scores to assign for each item using the information that is typically found in these files.

Internal Capacity items

Item 1 – Intact Cognitive Functioning

For this item you need to find the individuals assessed IQ. Every file will have this *somewhere*. First look in the Demographic Questionnaire and find Item 35 – Assessed IQ. Consult SAPROF-SO booklet and score accordingly.

If this is not recorded in the "Demographic Questionnaire", you can find their IQ score by looking for their WAIS-R score in the "Test Score Summary Sheet". If it is not there, find their WAIS-R scoring sheet among their psychometric scoring sheets.

You also need to check the "Presentation" section of their psychological report to see if there was anything about any impairments in cognitive functioning.

If they have a reported IQ at the higher end of the borderline range, and their psychologist reports their cognitive functioning was not a responsivity barrier (i.e., the individual was able to understand concepts and engage in treatment), consider a score of 1 (rather than 0). This guideline helps account for error in the estimation of full-scale IQ scores (especially given the abbreviated WAIS was used to estimate FSIQ).

Item 2 – Secure Attachment in Childhood

Here, you are looking for reports of good relationships with caregivers in childhood. For a score of 4, they need to be both: raised by a prosocial adult AND they need to have felt **loved and accepted**.

First, read the "Psychological Report" you located earlier. There is often a section about their childhood and connection to their caregivers. There will also be some relevant information in the "Demographic Questionnaire" and the handwritten document "Social History" among their therapy work.

Item 3 – Adaptive Schema

This is typically one of the hardest items to code.

First, comb through the "Psychological Report" looking for information related to:

1. Self-worth – grandiose or low.
2. Other people being trustworthy
3. The world being a safe place
4. The impact these opinions of themselves and others have on their lives.

There is a large amount of variability between reports in terms of their format. However, you are most likely to find information on these variables in the sections "Presentation", "Treatment Considerations", "Psychometric Assessment" and possibly "Treatment Provided". These are the sections concerned with information relating to recent functioning, which you are to prioritise over historical functioning.

Being a difficult group member can be taken as evidence of a maladaptive schema because this can be due to the individual feeling that others are overly critical or harmful.

Cognitive distortions and failing to take responsibility for their actions can **SOMETIMES** be seen as evidence of a maladaptive schema if it is obvious this is due to them having a sense of entitlement and having grandiose self-worth.

The presence of personality disorders can also be indicative of a maladaptive schema, in particular, borderline and narcissistic personality disorders.

You should also integrate information from psychometric scores. Get the "Test Score Summary Sheet" and look at whether their pre and post treatment scores on the following psychometrics meet significant levels.

1. Hostility Toward Women (HTW)
   This is can be indicative of whether they think women can be trusted.
2. University of California Loneliness Scale (UCLS)
   This can also be indicative of whether the individual believes others can be trusted and how safe the world is.
3. The Social Self-Esteem Inventory (SSEI)
   This can give an indication of how good they believe they are at interacting with others (Self-worth).
4. Hare Psychopathy Checklist (Hare PCL-R)
   When using this psychometric, DO NOT look their overall score. Instead find their scoring sheet among their psychometrics and see what their clinician rated them on the item "Grandiose Sense of Self-Worth" because these are indicative of having a maladaptive schema.

When considering the stability of the schema, it is useful to compare the pre and post treatment scores found in the "Test Score Summary Sheet". If someone has many clinically significant pre-treatment scores but protective post-treatment scores, the highest score you can give is a 2 – this is particularly important for the BDI scores.

Using all of the information, giving greater weight to recent functioning (i.e., the past 12 months), make a judgement call on what to score from analysing the scoring manual.

Item 4 – Empathy

It is important to remember when scoring this item, empathy for past victims is not relevant. Ideally, you are looking for evidence of empathetic behaviours while in prison.

First, look at their Hare PCL-R scoring sheet and see how they rated on the item "Lack of empathy". A score of 2 is indicative of a lack of empathy and a score of 0 means they have no lack of empathy. This score is rated by their main psychologist at the end of treatment. Use this as a "baseline" in your decision making. If they also have a score of 1 or 2 for "Grandiosity" this can also be seen as evidence of having a lack of empathy but should be considered with less weight.

If the person scores 1 or 2 on the Hare PCL-R Lack of Empathy Item, the highest score they can attain as a 1, regardless if they show empathy in two contexts. According to the manual, if someone shows empathy in two contexts but displays callous behaviour in another, they are assigned a maximum score of 1. The therapist would only assign a 1 or 2 on this Hare item if they had witnessed callous behaviour.

Then, go through the "Psychological Report" and there may something about how he is a "supportive group member" or how he helped others with their offence cycle etc. You can take this as evidence of empathetic behaviours present in **one context**.

If they are in a romantic relationship, you can consider this one context, but to qualify there needs to be behavioural evidence or specific comment on whether the person has showed empathy towards the partner. Vague comments from the therapist about how they are in a "supportive" or "caring" relationship will not suffice.

There may be some mention of their engagement in the "empathy module" of the programme. This module exclusively deals with victim empathy and is therefore to be **ignored** when scoring this item.

Next, you can ignore the 6-month suggested timeframe and look at their previous functioning before coming to prison. Involvement in charity groups, caring for a vulnerable family member or any other evidence of empathetic behaviours can be integrated into decision making but more weight should be given to recent functioning. You may find this information in the Presentence Report and current Report.

If the person was a supportive group member (1 of 2 Contexts), got a score of 0 on the PCL-R empathy item:
Assign them a 4 if there is evidence of pre-prison empathetic behaviours through involvement in altruistic groups like charities etc.
Assign them a 2 if there is historical evidence of a callous behaviour pattern (with no recent evidence), or if there is no evidence of current or historical empathic behaviours.
Assign them a 1 or lower if there is recent evidence of callous behaviours.

Use these rules to guide scoring in similar situations.

We want to make sure we are not just giving 4s to those who *don't present as callous.* We want to only give 4s to those who actually show empathy and engage in helpful behaviours.

Item 5 – Coping and Managing Stress

For this item, you are looking for evidence of *skills* that the individual has to manage stress and emotions. You are looking for the quantity of those skills *and* the perceived effectiveness of those skills in **reducing inner turmoil or stress**. You are also looking for evidence that they **actively pursue peace of mind** for example through a recreational activity or through mindfulness.

First, go to the psychological report and look for comment on these three things, in particular information from the mood management module under the Treatment section is useful.

Next, find the document "Rating Scale – Post Treatment" and see how they are scored on the item "Stress management skills" and consider this when allocating a score. Their score on the item "Problem solving skills" is also relevant but is not to be considered with the same weight.

Next, look at their STAXI – Anger suppression score on their "Test summary score sheet" and see if they score above the clinical threshold. If it is, this can be seen as poor coping. This is because if they

suppress their anger, they are not reducing their stress but rather, ignoring it. STAXI – Trait Anger and STAI – Trait Anxiety Scores are also relevant because if they are consistently angry and anxious, this can be seen as evidence of a lack of coping skills.

The Hare PCL-R Item "Poor Behavioural Controls" is also relevant here as the Hare criteria states that the person struggles to manage their moods resulting in behavioural incidents. This is of course also relevant to the Self-control item.

You should also look through the document "My Personal Statement" to see I there is relevant information here.

Integrate the information from these areas, giving greater weight to recent functioning, and decide on a score in consultation with the scoring manual.

Item 6 – Self-control and Managing Impulses

There is a natural relationship between the items "Coping" and "Self-Control". The Coping item is about the range and effectiveness of skills for regulating stress and emotions that can lead to problem behaviours, while Self-Control is about resisting temptation or managing impulses to engage in problem behaviour. Poor coping can lead to poor management of impulses so there is certainly some overlap. To illustrate the difference between the two, look at the following examples.

1. Becoming angry with someone and hitting them is consistent with someone showing **poor coping** and **poor self-control**.
2. Becoming angry and refraining from hitting someone but continuing to ruminate on the anger is consistent with someone showing **poor coping** but **good self-control.**
3. Becoming angry with someone but then calming themselves down using coping skills is consistent with **good coping.**
4. If someone is on a diet and has some unhealthy food (like chocolate cake), they are showing **poor self-control,** but this has little to do with coping, they simply wanted the cake. Note that this problem behaviour came about through a failure to manage impulses (self-control) not because of a failure to manage negative emotions (coping). However, if someone ate the chocolate cake as stress eating, this would indicate both poor coping and poor self-control.

Evidence of self-control can be gathered from various places throughout the file, but some information is to be weighted higher than others. For this reason, we have decided on a hierarchy.

The most important information to consider comes from comments made in the "Psychological Report". If someone has poor self-control, this may be commented on directly or will be evident through any mention of behavioural incidents in group. Some examples of incidents could be that they stormed out of group or were particularly aggressive or argumentative with other group members or the therapist.

The next best information to consider comes from psychometrics. Locate their Hare PCL-R scoring sheet and find what their scores were for the items "Poor behavioural controls" and "Impulsivity". Note that a score of 2 is indicative of poor controls and high impulsivity and a score of zero means they have adequate behavioural controls and impulsivity. Equally relevant information comes from their scores on the STAXI psychometric. If you look at their "Test Score Summary Sheet" you will see

they are given three STAXI scores and the scores next to them indicate when they are above or below threshold for being problematic. Of particular interest to this item are their scores on "Anger control" and "Anger expression". High "Anger suppression" is also relevant but can be seen as an indication of good self-control as oppose to bad coping per the last item.

The third most relevant content to integrate into your decision is evidence of previous functioning before coming to prison. Go through their "Demographic questionnaire" and "Psychological report" to look for a number of things present in their past that could indicate their level of self-control.

Good self-control is evident through:

1. High education level – having a university degree is evidence they have the self-control to stay on track. Note this is relative to their intellectual functioning.
2. Stable work history – this is mentioned directly in the "Demographic Questionnaire".
3. Other evidence of attaining long term goals.

Poor self-control is evident through:

1. History of violent outbursts.
2. History of drug or alcohol abuse – this is also commented on in the "Demographic questionnaire".

Scoring this item is a little different from the other items in that you can deviate from the specifics of the scoring manual.

Taking all relevant information into account, remembering to give greater weight to recent functioning, use the guide below to decide on a score:

4: Good self-control.
3: Above average self-control.
2: Average self-control.
1: Below average self-control.
0: Poor self-control.

Item 7 – Sexual Self-regulation

In coding this item, you are looking for evidence of the 4 criteria shown in the manual. **The scoring of this item deviates from how it is shown in the manual.** For each of the 4 criteria (or elements as they are called in the manual), you will mark whether they are Present (Y), Not Present (N) or Unclear/Omit (?) in each context box that you will divide into four (1 for each element). See the example below

Current		**Future**	
1. Y	2. N	1. Y	2. N
3. Y	4. ?	3. Y	4. ?

Note the scores are the same in each context because all current context scores are to be copied into the future context boxes for all "Internal capacity" and "Prosocial identity" items (like you have been doing).

For each element use their description near the top of the page as a guide and you can largely ignore the information in the *Scoring Classification and Examples* section.

For elements 1 and 2 you are looking for evidence of planning for release and general preparedness.

Element 1 - At the end of the report there will be a section called something like "Plans for release" or "Follow up plans and needs". You should be able to gauge whether their lifestyle avoids *their individual high-risk situations* (see end of report) and whether the author of the report believes this plan to be adequate. It is important to distinguish between a lifestyle that avoids high risk situations simply because of restraints present (e.g., conditions stipulating no contact with <16s) versus a lifestyle that avoids high risk situations for a better (and offence-free) life.

If the person is returning to a home with children present, you can still assign a Y if there are strict rules in place **AND** they are low risk.

Element 2 - In the report, the psychologist will sometimes make a comment about strategies in dealing with **inadvertently encountered high risk situations**. This is not always the case. In the "Treatment outcomes" section, there may be something about well they know their offence pattern, which is loosely related.

Some files (not many) may have a document bundled in with their work in therapy called "My personal statement" and there could be something about strategies here. There may also be another document called "Relapse Prevention Interview" where the individual is specifically asked to articulate their strategies for dealing with high risk situations and the clinician gives them a rating between 0 and 2, with 2 being good. This document is routinely missing.

This quantity of information required to make a sound judgement is often lacking for this element, and in these circumstances, you are to mark them as "?" (omit).

Elements 3 and 4 – These are usually hard to score because of a lack of information available about their sexual thoughts and habits. First, look in their psychological report to see how they responded to phallometric testing. Many are non-responders. There may also be some comment about their attraction to children in their report. There is also a handwritten document called "Sexual History" which can be found among the therapy work which may be useful if you can read it. Again, you may need to omit these if there is insufficient information.

Element 4 – Further information about element 4 can be taken from the "Post Treatment Rating Scale" item "Sexual Preoccupation".

Prosocial Identity items

Item 8 – Prosocial Sexual Interest

This item is about their sexual interests and whether there is evidence of pro-social or normal sexual interest. When considering whether they have deviant sexual interests, incest is irrelevant.

For this item it is useful to know whether there was a lengthy period between the offending taking place and them going to prison. This information can be found in the Presentence Report, The Sentencing Notes/Remarks and Summary of Facts.

As per the last item you will need information on their sexual preferences and habits. Specifically, you will want to know if:

1. If they have paedophilic or deviant sexual interests (rape fantasy etc.). Contrary to popular belief, just because someone has offended against a child, does not mean they are a paedophile. Look for patterns in their offending.

2. To see if they have had normal sexual relationships with an adult sexual partner both before and after the offending took place. If there is evidence that they had normal sexual relations in the 24 months before going to prison this will affect scoring. Do not give credit for having previous adult sexual relationships if there is evidence to suggest a pattern of entering adult relationships to gain access to children.

To find this information, first, check their results from phallometric testing. Next, read the full report. You can also find information on this from the "Social History: Adult" section of the Demographic Questionnaire. There will also be a handwritten document in the file called "Sexual History" and you may find some useful information here if you can read it. In deciding whether they have paedophilic interest, it will be useful to look at their number of previous victims. Also consider the age of the victims, if they were close to 16 years this is not indicative of paedophilic interest.

Scoring this one can be difficult so read the scoring manual carefully. Main parts to note are:

- If they have a history of deviant sexual interests (e.g., paedophilia) they cannot get a 4 unless there is evidence of 24 months of pro-social sexual interests in the 24 months between offending and coming to prison AND they have no current paedophilic interest.
- Some will not have a gap of two years between offending and coming to prison and the highest score they can receive is a 2 (bullet points 4 and 5 in the manual).

If there was at least two years between offending and coming to prison AND there is evidence of adult consenting sexual relations (or other evidence of arousal to adults) in that time, AND there is no evidence of having deviant interests in the last two years, assign them a 4.

If there is no evidence of appropriate sexual relations or arousal before OR after the offending, give them a 0. If they have current paedophilic interest and is no gap between them offending and coming to prison but has extensive experience with appropriate sexual relationships give them a 1.

Item 9 – Prosocial Sexual Identity

This one is easy if they scored **lower** than a 2 on the previous item. If they have, assign a 0. Otherwise, read on.

If the offending against children has occurred in the previous 12 months, the highest they can score is a 2. This information can be found in The Summary of Facts or Sentencing Notes.

If the individual has exclusively offended against female children and there is evidence of them being heterosexual (from information in the Demographic Questionnaire and handwritten "Sexual History" document) assign them a 4.

If the individual identifies as heterosexual but has more than one male victim, interpret this as possible evidence of hidden homosexual orientation and assign them a 0. Further evidence of this can come from if they never had a long-term relationship with a member of the opposite sex.

Item 10 – Goal-directed living

At the end of their most recent "Psychological Report", you will find information in their plans for release which should indicate their priorities and meaningful life goals. There is usually a comment on their search for a job or qualification and how "fruitful" their enquiries have been in setting this up for their release. If they have a career goals it is up to you to estimate how realistic these are.

You are also looking evidence of other types of goals for their release and evidence that they are actively working towards them. For example, these can be family, spiritual, or cultural goals. Consult the manual for more information on this.

As directed in the manual, evaluate the extent to which the goal is protective against further offending. More protective goals will provide structure to their lives, mastery, prosocial reward, satisfaction, etc.

To aid in your decision, you can go to their "Hare PCL-R" checklist (found among the psychometrics) and see how they scored on the item "Lack of realistic/long-term plans" and integrate this into your decision. Note that *not having a lack of realistic long-term plans* is not the same as *having pro-social and meaningful life goals* but this may give you some indication. Keep in mind that even men that have no discernible goals are routinely given 0 on this item.

If their goal is vague but possible and prosocial, assign a 1.
If they intend on enrolling in a course relevant to their goals, assign a 1.
If they have enrolled in a course relevant to their goals which is yet to start, assign a 2.

Item 11 – Motivation for Managing Risk

The manual stipulates that if they are rated as low assign them a rating of N/A. You will notice near the end of the report that they are given a risk rating. **You are to ignore this** because this is a rather subjective Clinician Rating and the manual is referring to risk assessed using an actuarial tool (a tool similar to this one that mathematically calculates risk-based on the presence of risk factors). You can however, integrate this into your decision making.

Next, if you are lucky, there will be a clinical opinion of what "stage" they are in at the end of their "Post/Post Treatment: Rating Scale" document i.e. Maintenance Stage. Only some of the later files will have this. If they are in the Maintenance stage both before and after treatment, give them a 3, otherwise, use a lower score.

No one can get a 4 because none of them have been in an uncontrolled environment post-treatment for 12 months.

Then go to the back of the main report and look for information relevant to managing risk. Usually, there will be something about:

a) Learning skills and strategies to lower risk.
b) Knowledge of their offence cycle.
c) Denial that they need to change.
d) General comment on their progress in the programme.

As a general rule, if they are still developing their risk management skills the highest they can get is a 2. If they have developed risk management skills and are applying them consider a 3.

Not all "Pre/Post Treatment Rating Scales" contain stage information. Some will only have the scales. Some of these scales are relevant to scoring this item. For example, motivation for change, cognitive distortions, perception of victim impact, empathy and stress management skills. The information you looked at for element 2 of Sexual Self-Regulation will also be relevant here.

Item 12 – Attitudes Towards Rules and Regulations

Start by going through their most recent "Psychological Report" looking for specific comment on their attitudes towards the rules of their incarceration. There may be something about their attitude towards the programme, the therapist, support network meetings, parole conditions, treatment goals or other rules and restrictions. There may also be more general and direct comments made by the clinician about how they respond to authority and rules. Some comments may be specific to the group treatment setting and some may reflect a long history of issues in these areas. Consider both the recency and persistence of these problems.

There may also some information that predates their imprisonment you can consider. For example having a history of *non-sexual offences* could be evidence of these issues, but more weight should be given to recent functioning. Be wary that this offending could also be due to a lack of self-control rather than a lack of respect for rules and authority. Similarly, they may have an unstable job history due to their issues with rules and bosses, but again, this could also be a sign of poor self-control.

There are some specific questions within the T.P.Q psychometric that are relevant to this item. Answers to questions 13 and 24 can be useful. It is important to remember that it is quite common that people will respond "True" to both of these and you should also consider that this psychometric is administered pre-treatment.

In scoring this item, it is reasonable to assume that if the individual holds attitudes against rules and authority, there would be evidence of this in the psychological report (either through reports of behaviour or direct comment). If no such evidence is found, it is reasonable to assume the individual

generally shows positive attitudes and compliance with rules and regulations and a **minimum score of 2 is indicated**.

Prosocial Connection items

Reminder: For these items, you are code the current and future contexts separately.

Item 13 – Work

For current context, score them a 2 if they show "consistent evidence of job preparedness activities", this refers to specific training for a job when they get out, which in my experience is rarely something that is available to them. You should be able to assess this from information on Post Release Plans at the end of the Psychological Report. For future context, assume they will get out soon and use this post release information. **If work is not confirmed or the plan seems vague, they cannot receive credit on this item.**

If they have a prison job assign a 2 for the current context.

If they are in the Kia Marama programme during or after 1998, assume they have a prison job and give them a 2 for the Current Context. Pre-1998 – if there's no mention of work or training, assume they don't have a prison job and assign a 0.

Item 14 – Leisure Activities

In scoring this item, *usually* the Psychological Report will be the only relevant document.

For the current context, if recreational activities are not mentioned in the Psychological Report, you can omit this item rather than give them a zero. This is because recreational activities while in prison were not usually commented on during this era, so you cannot assume they are not present from their lack of inclusion.

For current context, do not assume other group members or fellow prisoners are prosocial peers unless otherwise stated. Instead, assume they are a mix of antisocial and prosocial peers.

For the future context, it is impossible to know whether a particular activity will be with prosocial peers or not, so you can largely ignore this part of scoring this item.

In deciding how protective their leisure activities will be, consider if they have previous experience doing that activity and how much they enjoy it. Higher scores should be given to individuals whom you can estimate will stick with their leisure activity upon release. Therefore, if it is known they like a particular activity they plan to engage in, it should be more protective than a plan to try something new. Further, vague plans to try a new hobby they have not researched would receive a lower score, like a 1. If they name a leisure activity that sounds obscure or uncommon (apart from bone carving and glass-painting - hobbies available in Kia Marama), this can be seen as an indication that it is likely they have done it before so it is therefore more likely to be protective.

An amendment to the scoring manual is that you should give a 3 if activities they plan engage in are regular and prosocial but require no interaction with other people. For example, church or the gym.

Item 15 & 16 – Social Network and Emotional Connection to Adults

The item Social Network deals more with the net effect of relationships (protective vs risky) and Emotional Connection deals more with the quality or nature of individual (and non-romantic) relationships.

When scoring Emotional Connection to Adults it is important to remember intimate partner relationships do not count (as stated in the manual).

Support Network: At the end of their most recent psychological report, there will be some information about their support network and a plan for a meeting where they will discuss their Relapse Prevention Plan with this group. Sometimes, at the front, there will be a list of "Follow up support persons".

If someone is labelled in the front of the report as a support person, they have agreed to this, so you can assume their status in the network is legitimate. Whether or not you believe the individual actually feels comfortable calling on these people in times of need or how prosocial these people are is up to your discretion based on information available (note: antisocial influences are unlikely to be named as support people given therapy staff have some input into support network planning).

A Reverend or Priest is not classed as a friend or volunteer, so they are classed as a separate system.

There may also be a document labelled "Post release plans" that can contain some relevant information but this is completed months before the final report so this information should only be used if it is not available elsewhere.

Unless you have information to the contrary, Social Network scores will be the same in both the Current and Future Contexts.

Emotional Connection to Adults: For this item you can consider a reciprocal relationship with a fellow group member as a prosocial relationship. You may note this is different to item 14. It is also important to remember this relationship will end when the person is released (Future Context).

There is sometimes insufficient information about reciprocal relationships, so it is not uncommon to omit Emotion Connection to Adults.

Item 17 – Intimate Relationship

Note the highest score you can give for the current context is a 2. For future context you can look at information on this in the psychological report and there are indications in the demographic questionnaire.

Part of scoring this item is assessing whether, upon release, they will be having normative (refer to manual for clarifications) sex with their partner. To best estimate this look at:

a. The level of satisfaction they claim to have in their current relationship from the Demographic Questionnaire.

b. The level of support the partner has for the individual gathered from psychologist comment in their report.
c. Comments made in their handwritten "Sexual History" document, found among their course work at the rear of their file.

You will also have to make a judgement about whether their partner is prosocial. You can assume this unless there is evidence to the contrary in the psychological report and presentence report.

Stability items

Item 18 – Housing Stability

For current context all will receive a 4.

Information for coding the future context can be found in "Post Release Plans" or a similar section at the end of their psychological report.

Item 19 – Financial Stability

For this item you are looking for information on their:

1. Savings and assets. This information can often be found in their Pre-sentence Report. You can also use this information to gauge their saving ability. There may also be information on their general financial situation in reports from previous probation officers which is usually found in the file near the pre-sentence report.

As a general rule, someone can receive a 4 on this item if they are in a position to cover at least one month's living expense and the **ability** to manage unexpected expenses. This ability should be evident from a history of sound financial management (savings, no previous bankruptcies etc.). Having a history of merely making ends meet and never "getting ahead" would indicate a maximum score of 2.

2. Post-release employment situation. See "Post-Release Plans" and other information you used to code the Work item.

If someone has no debts or savings or confirmed employment, but owns a home, assign a 3 for future context.

It is highly unlikely there will be any comment on their ability to save while in prison.

Professionally Provided Support items

Item 20 – Sex Offence Specific Treatment

For this item all will receive a 4 for the current context because in their *current context* they have just finished the Kia Marama programme.

For the future context, information on their treatment plans after release should be available at the end of the Psychological Report.

If they will be receiving counselling for something that was a factor their offending (drugs or mental health) upon their release, you can count this as sex offence specific treatment.

Men who will attend Relapse Prevention meetings and Kia Marama support groups receive a 4. If this is not available but they will be seeing a psychologist, assign a 4.

Every person in this sample will benefit from some kind of psychological intervention, therefore do not assign N/R.

Item 21 – Medication

This one will almost always be N/A. If there is no mention of medication in their report it is safe to assume they are not on any. I have not come across a file yet where someone was prescribed medication to reduce their sexual drive.

If they are medicated for a mental illness, addiction or physical health problem that was a factor in their original offending, this medication counts as being protective and you should score this item.

If you find they are on some kind of medication and you do not know if it affects libido, or if the underlying health problem was a factor in their offending, score them N/A.

Item 22 – Perception if the Therapeutic/Supervisory Alliance

It is incredibly uncommon for reports from this era to make any comment about a therapeutic alliance, therefore do not rate this item.

Item 23 – Supervised Living

For current context, all will receive a 4 because they are in prison.

For future context, there should be enough information about who they will be living with to score this item at the end of their Psychological Report. If this person is labelled as a support person, you can assume they will be knowledgeable of the risk management plan if they are invited to attend the release prevention meeting. Unless commented on, you can assume the support person is prosocial. If you believe support person is knowledgeable of the risk management plan, would intervene if needed, and there is no evidence against them being prosocial, assign a 2. If you find evidence that the support person may not be prosocial, may not be made aware of the risk management plan or may not intervene if needed, consider a lower score.

Not all reports label who will attend the Relapse Prevention Meeting. In these situations, you can assume the support person who will be living with the individual will be aware of the risk management plan.

If the person will be living alone but there will be a property manager living on site who is aware of their offending, assign a 1.

Scoring manual amendment: Give the person a 3 if they live alone but are subject to meetings with their parole officer every day or every second day.

Item 24 – External Control

For current context, all will receive a 4 because they are in prison.

For future context, use information about their release conditions at the end of their psychological report. If there is a lack of information around their parole conditions, assume they are subject to visits on a weekly or fortnightly basis and assign a 2.

Congratulations! You are done

Appendix C

Detailed Results of Interrater Reliability Analyses

Table C1

Interrater Reliability of SAPROF-SO Scores for Current Context Ratings

SAPROF-SO items and domains	ICC *n*	ICC	% Agree *n*	% Agree
Personal				
1. Intact cognitive functioning	22	.94 [.86,.97]	22	95.5
2. Secure attachment in childhood	21	.81 [.60,.92]	21	90.5
3. Adaptive schema	22	.72 [.43,.88]	22	95.5
4. Empathy	19	.81 [.58,.92]	19	94.7
5. Coping	22	.69 [.39,.86]	22	95.5
6. Self-control	22	.82 [.56,.93]	22	95.5
7. Sexual self-regulation	[a]	---	3	66.7
8. Prosocial sexual interests	21	.88 [.73,.95]	21	95.2
9. Prosocial sexual identity	22	.75 [.50,.89]	22	90.9
10. Goal-directed living	21	.62 [.26,.83]	21	95.2
11. Motivation for managing risk	22	.89 [.75,.95]	22	100
12. Attitudes towards rules and regulations	22	.83 [.59,.93]	22	90.9
13. Work	21	.89 [.76,.96]	21	95.2
14. Leisure activities	[a]	---		
15. Social network	22	.74 [.46,.88]	22	95.5
16. Emotional connection to adults	15	.52 [.01,.81]	15	86.7
17. Intimate relationship	22	.94 [.86,.98]	22	100
18. Housing stability	[b]	---	22	100
19. Financial management	21	.90 [.77,.96]	21	100
Personal (Current) subtotal	22	.98 [.93,.99]		
Professionally Provided				
20. Sexual offence-specific treatment	[b]	---	22	100
21. Medication	[a]	---	1	100
23. Supervised living	[b]	---	22	100
24. External control	[b]	---	22	100
Professionally Provided (Current) subtotal	[b]	---		
Total (Current) score	22	.98 [.93,.99]		

Note. SAPROF-SO = Structured Assessment of PROtective Factors – Sexual Offence version (Willis et al., 2017-2019). ICC = Intraclass correlation coefficient. Cases were included in item analyses when neither rater omitted the item. Missing scores were replaced with case mean domain scores in calculation of subtotal and total scores. Not applicable ratings were replaced with scores of 0. Sexual Self-Regulation scores were deemed valid when zero elements were omitted.

[a] Could not be computed due to lack of valid scores.

[b] Could not be computed due to no variance in scores.

Table C2

Interrater Reliability of SAPROF-SO Scores for Future Context Ratings

SAPROF-SO items and domains	ICC *n*	ICC	% Agree *n*	% Agree
Personal				
13. Work	21	.98 [.95,.99]	21	100
14. Leisure activities	20	.92 [.81,.97]	20	100
15. Social network	22	.77 [.53,.90]	22	95.5
16. Emotional connection to adults	16	.67 [.27,.87]	16	93.8
17. Intimate relationship	22	.93 [.83,.97]	22	95.5
18. Housing stability	21	.98 [.94.99]	21	100
19. Financial management	20	.78 [.53,.91]	20	100
Personal (Future) subtotal	22	.98 [.95,.99]		
Professionally provided				
20. Sexual offence-specific treatment	20	.60 [.23,.82]	20	90
21. Medication	[a]	---	1	100
22. Supervised living	21	.48 [.09,.75]	21	90.5
23. External control	[b]	---	21	100
Professionally provided (Future) subtotal	22	.90 [.77,.96]		
Total (Future) score	22	.98 [.95,.99]		

Note. SAPROF-SO = Structured Assessment of PROtective Factors – Sexual Offence version (Willis et al., 2017-2019). ICC = Intraclass correlation coefficient. Cases were included in item analyses when neither rater omitted the item. Future context statistics for the first 12 items are identical to those presented in Table B1 as Current context scores were used as Future context scores in the present study. Missing scores were replaced with case mean domain scores in calculation of subtotal and total scores. Not applicable and not required ratings were replaced with scores of 0. Sexual Self-Regulation scores were deemed valid when zero elements were omitted.

[a] Could not be computed due to lack of valid scores.

[b] Could not be computed due to no variance.